

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

G.729.1

Amendment 4

(06/2008)

SERIES G: TRANSMISSION SYSTEMS AND MEDIA,
DIGITAL SYSTEMS AND NETWORKS

Digital terminal equipments – Coding of analogue signals
by methods other than PCM

G.729-based embedded variable bit-rate coder: An
8-32 kbit/s scalable wideband coder bitstream
interoperable with G.729

**Amendment 4: New Annex C (DTX/CNG scheme)
plus corrections to main body and Annex B**

Recommendation ITU-T G.729.1 (2006) –
Amendment 4



International
Telecommunication
Union
Documents

Document Sharing Hub

ITU-T G-SERIES RECOMMENDATIONS
TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS

INTERNATIONAL TELEPHONE CONNECTIONS AND CIRCUITS	G.100–G.199
GENERAL CHARACTERISTICS COMMON TO ALL ANALOGUE CARRIER-TRANSMISSION SYSTEMS	G.200–G.299
INDIVIDUAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON METALLIC LINES	G.300–G.399
GENERAL CHARACTERISTICS OF INTERNATIONAL CARRIER TELEPHONE SYSTEMS ON RADIO-RELAY OR SATELLITE LINKS AND INTERCONNECTION WITH METALLIC LINES	G.400–G.449
COORDINATION OF RADIOTELEPHONY AND LINE TELEPHONY	G.450–G.499
TRANSMISSION MEDIA AND OPTICAL SYSTEMS CHARACTERISTICS	G.600–G.699
DIGITAL TERMINAL EQUIPMENTS	G.700–G.799
General	G.700–G.709
Coding of analogue signals by pulse code modulation	G.710–G.719
Coding of analogue signals by methods other than PCM	G.720–G.729
Principal characteristics of primary multiplex equipment	G.730–G.739
Principal characteristics of second order multiplex equipment	G.740–G.749
Principal characteristics of higher order multiplex equipment	G.750–G.759
Principal characteristics of transcoder and digital multiplication equipment	G.760–G.769
Operations, administration and maintenance features of transmission equipment	G.770–G.779
Principal characteristics of multiplexing equipment for the synchronous digital hierarchy	G.780–G.789
Other terminal equipment	G.790–G.799
DIGITAL NETWORKS	G.800–G.899
DIGITAL SECTIONS AND DIGITAL LINE SYSTEM	G.900–G.999
QUALITY OF SERVICE AND PERFORMANCE – GENERIC AND USER-RELATED ASPECTS	G.1000–G.1999
TRANSMISSION MEDIA CHARACTERISTICS	G.6000–G.6999
DATA OVER TRANSPORT – GENERIC ASPECTS	G.7000–G.7999
PACKET OVER TRANSPORT ASPECTS	G.8000–G.8999
ACCESS NETWORKS	G.9000–G.9999

For further details, please refer to the list of ITU-T Recommendations.

Recommendation ITU-T G.729.1

G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729

Amendment 4

New Annex C (DTX/CNG scheme) plus corrections to main body and Annex B

Summary

This Recommendation describes an 8-32 kbit/s scalable wideband speech and audio coding algorithm interoperable with G.729, G.729A and G.729B.

The output of the G.729EV coder has a bandwidth of 50-4000 Hz at 8 and 12 kbit/s and 50-7000 Hz from 14 to 32 kbit/s. At 8 kbit/s, G.729EV is fully interoperable with G.729, Annex A/G.729 and Annex B/G.729. Hence, an efficient deployment in existing G.729-based VoIP infrastructures is foreseen. The coder operates on 20 ms frames and has an algorithmic delay of 48.9375 ms. By default, the encoder input and decoder output are sampled at 16 kHz.

The encoder produces an embedded bitstream structured in 12 layers corresponding to 12 available bit rates from 8 to 32 kbit/s. The bitstream can be truncated at the decoder side or by any component of the communication system to adjust "on the fly" the bit rate to the desired value with no need for outband signalling.

The underlying algorithm is based on a three-stage coding structure: embedded Code-Excited Linear Prediction (CELP) coding of the lower band (50-4000 Hz), parametric coding of the higher band (4000-7000 Hz) by Time-Domain Bandwidth Extension (TDBWE), and enhancement of the full band (50-7000 Hz) by a predictive transform coding technique referred to as Time-Domain Aliasing Cancellation (TDAC).

Amendment 1 introduces the new Annex A containing the RTP payload format, capability identifiers and parameters for signalling of G.729.1 capabilities using H.245. Both format and capability parameters are fully compatible with the corresponding G.729.1 RTP definitions to allow seamless interoperability. Besides the new Annex, Amendment 1 to G.729.1 incorporates changes needed to correct defects in G.729.1 and provides new, more comprehensive test vectors.

Amendment 2 introduces the new Annex B, which defines an alternative implementation of the G.729.1 algorithm using floating point arithmetic to be used for implementation on DSP hardware optimized for floating-point operations. The accompanying floating point C-code is fully interoperable with the fixed-point C-code.

Amendment 3 extends the low-delay functionality of main body and Annex B to the first wideband bit rate (14 kbit/s). It also incorporates changes needed to correct defects in the text and C-code of G.729.1 main body and Annex B.

Amendment 4 introduces a new Annex C specifying a discontinuous transmission (DTX) and comfort noise generation for G.729.1. With this annex, the G.729.1 encoder is capable of generating a silence insertion description (SID) each time an update of the ambient background noise parameters is required to maintain the quality of the generated background noise. The SID information includes a core lower band layer, which can be decoded by the decoder of Annex B/G.729, an enhancement lower band layer and a higher band layer. The non-transmission between SID updates and the small size of the SID provide a significant reduction of bandwidth during inactive segments. Besides this new annex, Amendment 4 incorporates changes needed to correct defects identified in G.729.1 C source code (main body and Annex B), provides a revised set of test vectors, and updates the complexity figures table of G.729.1 text.

Source

Amendment 4 to Recommendation ITU-T G.729.1 (2006) was approved on 29 June 2008 by ITU-T Study Group 16 (2005-2008) under the Recommendation ITU-T A.8 procedure.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure e.g. interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2009

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

CONTENTS

Page

1	Scope	1
2	References.....	1
3	Abbreviations.....	2
4	Notations and conventions.....	2
5	General description of the coder.....	5
5.1	Encoder.....	5
5.2	Decoder.....	6
5.3	Frame erasure concealment (FEC)	8
5.4	Coder modes.....	8
5.5	Bit allocation to coder parameters and bitstream layer format.....	9
5.6	Algorithmic delay	11
5.7	Computational complexity and storage requirement.....	11
5.8	Coder description.....	11
6	Functional description of the encoder.....	11
6.1	QMF analysis filterbank, signal downscaling and spectral folding of the higher band	11
6.2	Pre-processing of lower band	13
6.3	Narrowband embedded CELP encoder (Layers 1 and 2).....	13
6.4	Pre-processing of the higher band	24
6.5	TDBWE encoder (Layer 3)	24
6.6	TDAC encoder (Layers 4 to 12).....	29
7	Functional description of the decoder.....	40
7.1	Narrowband embedded CELP decoder (Layers 1 and 2).....	40
7.2	TDBWE decoder (Layer 3)	41
7.3	TDAC decoder (Layers 4 to 12).....	49
7.4	Post-filtering of the lower band.....	57
7.5	Spectral folding of higher band, signal upscaling and QMF synthesis filterbank.....	58
7.6	Frame erasure concealment	60
7.7	Bit-rate switching	73
8	Description of the transmitted parameter indices	75
9	Bit-exact description of the G.729EV coder.....	77
9.1	Use of the simulation software	77
9.2	Organization of the simulation software	78



Documents

Document Sharing Hub

Rec. ITU-T G.729.1 (2006)/Amd.4 (06/2008)

iii

	Page
Annex A – Packet format, capability identifiers and capability parameters.....	85
A.1 References	85
A.2 Packet structure for G.729.1 frames	85
A.3 Capability identifiers and parameters for use with ITU-T Rec. H.245	85
A.4 Interoperability with ITU-T Rec. G.729.....	87
Annex B – Reference floating-point implementation for G.729.1	88
B.1 Scope	88
B.2 References	88
B.3 Overview	88
B.4 Algorithmic description.....	88
B.5 ANSI C Code.....	88
Annex C – DTX/CNG scheme	91
C.1 Scope	91
C.2 References	91
C.3 Definitions	91
C.4 Abbreviations and acronyms	91
C.5 Conventions.....	91
C.6 General description of the DTX, SID and CNG components	96
C.7 Functional description of the silence compression encoder	100
C.8 Functional description of the silence compression decoder	107
C.9 Memory update.....	114
C.10 Description of the transmitted parameter indices.....	114
C.11 Bit-exact description of the silence compression of G.729.1	115

Electronic attachment – C source code and test vectors

.....

ITU-T Recommendation G.729.1

G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729

Amendment 4

New Annex C (DTX/CNG scheme) plus corrections to main body and Annex B

1 Scope

This Recommendation contains the description of an algorithm extending ITU-T Rec. G.729 for the scalable coding of narrowband and wideband speech and audio signals at 8-32 kbit/s.

This Recommendation is organized as follows. The references, abbreviations and notations used throughout this Recommendation are defined in clauses 2, 3 and 4, respectively. Clause 5 gives a general outline of the G.729EV algorithm. The G.729EV encoder and decoder principles are discussed in clauses 6 and 7, respectively. The transmitted parameters are presented in clause 8. Clause 9 describes the software that defines this coder in 16-32 bits fixed-point arithmetic.

2 References

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- ITU-T Recommendation G.191 (2005), *Software tools for speech and audio coding standardization*.
- ITU-T Recommendation G.192 (1996), *A common digital parallel interface for speech standardization activities*.
- ITU-T Recommendation G.729 (1996), *Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)*.
- ITU-T Recommendation G.729 – Annex A (1996), *Reduced complexity 8 kbit/s CS-ACELP speech codec*.
- ITU-T Recommendation G.729 – Annex B (1996), *A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70*.

3 Abbreviations

The acronyms used in this Recommendation are summarized in Table 1.

Table 1 – Glossary of acronyms

Acronym	Description
ACELP	Algebraic CELP
BWE	BandWidth Extension
CELP	Code-Excited Linear Prediction
CNG	Comfort Noise Generator
DEMUX	DEMultipleXer
DTX	Discontinuous Transmission
FEC	Frame Erasure Concealment
FIR	Finite Impulse Response
FFT	Fast Fourier Transform
G.729EV	G.729 Embedded Variable bit rate
HB	Higher Band
HPF	High Pass Filter
IIR	Infinite Impulse Response
IP	Internet Protocol
LB	Lower-Band
LP	Linear Prediction
LPF	Low Pass Filter
LSB	Least Significant Bit
LSF	Line Spectrum Frequency
LSP	Line Spectrum Pair
LTP	Long-Term Prediction
MDCT	Modified Discrete Cosine Transform
MSB	Most Significant Bit
MUX	MUltipleXer
PCM	Pulse Code Modulation
QMF	Quadrature Mirror Filterbank
SID	Silence Insertion Descriptor
TDAC	Time-Domain Aliasing Cancellation
TDBWE	Time-Domain BandWidth Extension
VoIP	Voice over IP
VQ	Vector Quantization
WB	Wideband
WMOPS	Weighted Million Operations Per Second

4 Notations and conventions

To be consistent with the text of ITU-T Rec. G.729, the 20 ms frames used by G.729EV will be referred to as *superframes*, whereas the 10 ms frames and the 5 ms subframes involved in the CELP processing will be respectively called *frames* and *subframes*.

Throughout this Recommendation, the G.729 notations are reused. The notational conventions are detailed below:

- Codebooks are denoted by calligraphic characters (e.g., $\mathcal{C}$).
- Time-domain signals are denoted by their symbol and a sample index between parenthesis, e.g., $s(n)$. The variable n is used as sample index.
- Frequency-domain transforms are denoted by converting the related time-domain signal to capital letters, e.g., $S(k)$ is the transform of $s(n)$. The variable k is used as coefficient index.
- Superscript indices between parentheses (e.g., $g^{(m)}$) are used to indicate time-dependency of variables. The variable m refers, depending on the context, to either a frame or subframe index, and the variable n to a sample index.
- Recursion indices are identified by a superscript between square brackets (e.g., $E^{[k]}$).
- Subscripts indices identify a particular element in a coefficient array.
- The symbol $\hat{}$ identifies a quantized version of a parameter (e.g., $\hat{g}_c$).
- Parameter ranges are given between square brackets, and include the boundaries (e.g., $[0.6, 0.9]$).
- The function $\text{int}()$ denotes truncation to its integer value.
- The function $\text{even}()$ returns 1 if the argument is an even integer number and 0 otherwise.
- The function $\text{round}()$ denotes the rounding to the nearest integer.
- The floating-point numbers used are rounded versions of the values used in the 16-bit fixed-point ANSI C implementation.

Table 2 lists the most relevant symbols used throughout this Recommendation.

Table 2 – Glossary of most relevant symbols

Type	Name	Description
Filters	$1/\hat{A}(z)$	Quantized LP synthesis filter
	$H_{hl}(z)$	Lower-band high-pass filter
	$H_p(z)$	Long-term postfilter
	$H_s(z)$	Short-term postfilter
	$H_t(z)$	Tilt-compensation filter
	$H_{h2}(z)$	Higher-band low-pass filter
	$H_l(z)$	QMF low pass analysis filter
	$H_2(z)$	QMF high pass analysis filter
	$G_l(z)$	QMF low pass synthesis filter
	$G_2(z)$	QMF high pass synthesis filter
	$P(z)$	Pre-filter for fixed codebook in G.729
	$W(z)$	Weighting filter in embedded CELP encoder
	$W_{LB}(z)$	Lower-band difference weighting filter
Signals	$c'(n)$	12 kbit/s layer fixed-codebook codevector
	$s_{WB}(n)$	Wideband input signal
	$s_{LB}^{qmf}(n)$	QMF low pass analysis filter output signal after decimation

Table 2 – Glossary of most relevant symbols

Type	Name	Description
	$s_{LB}(n) = s(n)$	Pre-processed lower band signal
	$d_{LB}(n)$	Lower-band difference signal between $s(n)$ and $\hat{s}_{enh}(n)$
	$d_{LB}^w(n)$	Lower-band difference weighting filter output signal
	$s_{HB}^{fold}(n)$	QMF analysis filter output signal after decimation and spectral folding
	$s_{HB}(n)$	Pre-processed higher band signal
	$\hat{s}(n)$	8 kbit/s CELP decoded signal
	$\hat{s}_{enh}(n)$	12 kbit/s CELP decoded signal
	$\hat{s}_{LB}(n)$	Lower-band decoded signal before postfiltering
	$\hat{s}_{LB}^{post}(n)$	Postfiltered lower-band decoder signal
	$\hat{s}_{LB}^{hpf}(n)$	High-pass filtered postfiltered lower band reconstructed signal
	$\hat{s}_{LB}^{qmf}(n)$	Decoded lower-band signal (input to QMF synthesis filterbank)
	$\hat{s}_{HB}^{qmf}(n)$	Decoded higher-band signal (input to QMF synthesis filterbank)
	$\hat{s}_{HB}^{bwe}(n)$	TDBWE decoded higher band signal
	$\hat{s}_{HB}(n)$	Decoded higher-band signal
	$\hat{d}_{LB}^w(n)$	Decoded lower-band weighted difference signal
	$\hat{d}_{LB}(n)$	Decoded lower-band difference signal
	$\hat{s}_{HB}^{fold}(n)$	Decoded folded higher band signal
	$D_{LB}^w(k)$	MDCT spectrum of $d_{LB}^w(n)$
	$S_{HB}(k)$	MDCT spectrum of $s_{HB}(n)$
	$\hat{D}_{LB}^w(k)$	MDCT spectrum of $\hat{d}_{LB}^w(n)$
	$\hat{S}_{HB}^{bwe}(k)$	MDCT spectrum of $\hat{s}_{HB}^{bwe}(n)$
	$\hat{S}_{HB}(k)$	MDCT spectrum of $\hat{s}_{HB}(n)$
Parameters	$\hat{a}_i$	Quantized lower-band LP coefficients
	α_{enh}	Tri-pulse parameter
	α	FEC attenuation factor
	τ	Position of the last glottal pulse

5 General description of the coder

The G.729EV coder is an 8-32 kbit/s scalable wideband (50-7000 Hz) extension of ITU-T Rec. G.729. By default, the encoder input and decoder output are sampled at 16 000 Hz. The bitstream produced by the encoder is scalable and consists of 12 embedded layers, which will be referred to as Layers 1 to 12. Layer 1 is the *core layer* corresponding to a bit rate of 8 kbit/s. This layer is compliant with G.729 bitstream, which makes G.729EV interoperable with G.729. Layer 2 is a narrowband enhancement layer adding 4 kbit/s, while Layers 3 to 12 are wideband enhancement layers adding 20 kbit/s with steps of 2 kbit/s.

This coder is designed to operate with a digital signal sampled at 16000 Hz followed by conversion to 16-bit linear PCM for the input to the encoder. However, the 8000 Hz input sampling frequency is also supported. Similarly, the format of the decoder output is 16-bit linear PCM with a sampling frequency of 8000 or 16000 Hz. Other input/output characteristics should be converted to 16-bit linear PCM with 8000 or 16000 Hz sampling before encoding, or from 16-bit linear PCM to the appropriate format after decoding. The bitstream from the encoder to the decoder is defined within this Recommendation.

The G.729EV coder is built upon a three-stage structure: embedded Code-Excited Linear-Prediction (CELP) coding, Time-Domain Bandwidth Extension (TDBWE) and predictive transform coding that will be referred to as Time-Domain Aliasing Cancellation (TDAC). The embedded CELP stage generates Layers 1 and 2 which yield a narrowband synthesis (50-4000 Hz) at 8 and 12 kbit/s. The TDBWE stage generates Layer 3 and allows producing a wideband output (50-7000 Hz) at 14 kbit/s. The TDAC stage operates in the Modified Discrete Cosine Transform (MDCT) domain and generates Layers 4 to 12 to improve quality from 14 to 32 kbit/s. TDAC coding represents jointly the weighted CELP coding error signal in the 50-4000 Hz band and the input signal in the 4000-7000 Hz band.

The G.729EV coder operates on 20 ms frames. However, the embedded CELP coding stage operates on 10 ms frames, like G.729. As a result two 10 ms CELP frames are processed per 20 ms frame. In the following, to be consistent with the text of ITU-T Rec. G.729, the 20 ms frames used by G.729EV will be referred to as *superframes*, whereas the 10 ms frames and the 5 ms subframes involved in the CELP processing will be respectively called *frames* and *subframes*.

Clauses 5.1 and 5.2 present a general description of the encoder and the decoder parts, while Clause 5.3 addresses the specific case of frame erasure concealment. The G.729EV coder has a highly flexible structure, therefore several modes are allowed at the encoder side as well as at the decoder side; the different modes are described in 5.4. The bit allocation, delay and complexity are detailed in clauses 5.5 and 5.6.

5.1 Encoder

A functional diagram of the encoder part is presented in Figure 1. The encoder operates on 20 ms input superframes. By default, the input signal $s_{WB}(n)$ is sampled at 16000 Hz. Therefore, the input superframes are 320 samples long.

The input signal $s_{WB}(n)$ is first split into two sub-bands using a QMF filter bank defined by the filters $H_1(z)$ and $H_2(z)$. The lower-band input signal $S_{LB}^{qmf}(n)$ obtained after decimation is pre-processed by a high-pass filter $H_{hl}(z)$ with 50 Hz cut-off frequency. The resulting signal $s_{LB}(n)$ is coded by the 8-12 kbit/s narrowband embedded CELP encoder. To be consistent with ITU-T Rec. G.729, the signal $S_{LB}(n)$ will also be denoted $s(n)$. The difference $d_{LB}(n)$ between $s(n)$ and the local synthesis $\hat{s}_{enh}(n)$ of the CELP encoder at 12 kbit/s is processed by the perceptual weighting filter $W_{LB}(z)$. The parameters of $W_{LB}(z)$ are derived from the quantized LP coefficients of the CELP encoder. Furthermore, the filter $W_{LB}(z)$ includes a gain compensation which

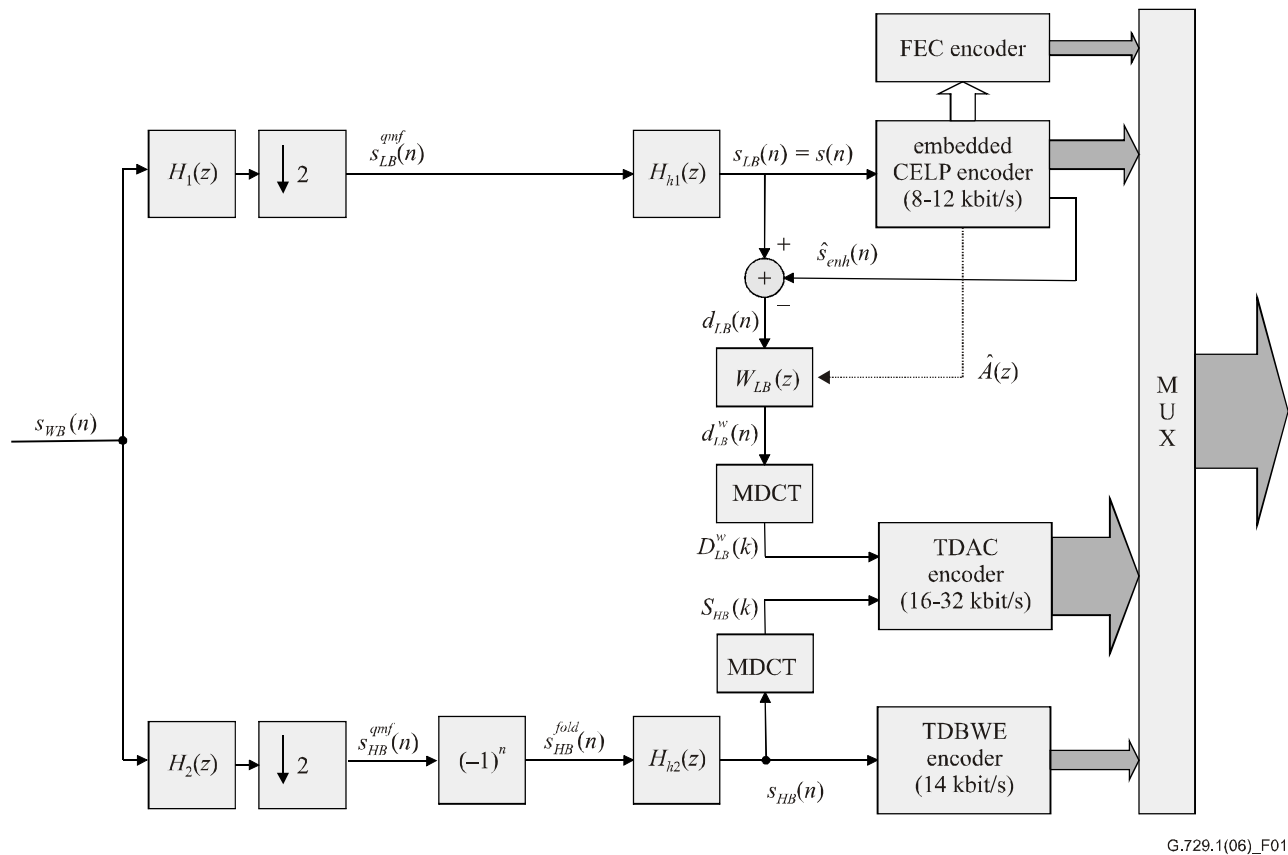


guarantees the spectral continuity between the output $d_{LB}^w(n)$ of $W_{LB}(z)$ and the higher-band input signal $s_{HB}(n)$. The weighted difference $d_{LB}^w(n)$ is then transformed into frequency domain by MDCT.

The higher-band input signal $s_{HB}^{fold}(n)$ obtained after decimation and spectral folding by $(-1)^n$ is pre-processed by a low-pass filter $H_{h2}(z)$ with 3000 Hz cut-off frequency. The resulting signal $s_{HB}(n)$ is coded by the TDBWE encoder. The signal $s_{HB}(n)$ is also transformed into frequency domain by MDCT.

The two sets of MDCT coefficients, $D_{LB}^w(k)$ and $S_{HB}(k)$, are finally coded by the TDAC encoder.

In addition, some parameters are transmitted by the frame erasure concealment (FEC) encoder in order to introduce parameter-level redundancy in the bitstream. This redundancy allows improving quality in the presence of erased superframes.



G.729.1(06)_F01

Figure 1 – High-level block diagram of the encoder

5.2 Decoder

A functional diagram of the decoder is presented in Figure 2. The specific case of frame erasure concealment is not considered in this figure and will be presented in 5.3. The decoding depends on the actual number of received layers or equivalently on the received bit rate.

If the received bit rate is:

- 8 kbit/s (Layer 1): The core layer is decoded by the embedded CELP decoder to obtain $\hat{s}_{LB}(n) = \hat{s}(n)$. Then $\hat{s}_{LB}(n)$ is postfiltered into $\hat{s}_{LB}^{post}(n)$ and post-processed by a high-pass filter (HPF) into $\hat{s}_{LB}^{qmf}(n) = \hat{s}_{LB}^{hpf}(n)$. The QMF synthesis filterbank defined by the filters $G_1(z)$ and $G_2(z)$ generates the output with a high-frequency synthesis $\hat{s}_{HB}^{qmf}(n)$ set to zero.

- 12 kbit/s (Layers 1 and 2): The core layer and narrowband enhancement layer are decoded by the embedded CELP decoder to obtain $\hat{s}_{LB}(n) = \hat{s}_{enh}(n)$, and $\hat{s}_{LB}(n)$ is then postfiltered into $\hat{s}_{LB}^{post}(n)$ and high-pass filtered to obtain $\hat{s}_{LB}^{qmf}(n) = \hat{s}_{LB}^{hpf}(n)$. The QMF synthesis filterbank generates the output with a high-frequency synthesis $\hat{s}_{HB}^{qmf}(n)$ set to zero.
- 14 kbit/s (Layers 1 to 3): In addition to the narrowband CELP decoding and lower-band adaptive postfiltering, the TDBWE decoder produces a high-frequency synthesis $\hat{s}_{HB}^{bwe}(n)$ which is then transformed into frequency domain by MDCT so as to zero the frequency band above 3000 Hz in the higher-band spectrum $\hat{S}_{HB}^{bwe}(k)$. The resulting spectrum $\hat{S}_{HB}(k)$ is transformed in time domain by inverse MDCT and overlap-add before spectral folding by $(-1)^n$. In the QMF synthesis filterbank the reconstructed higher band signal $\hat{s}_{HB}^{qmf}(n)$ is combined with the respective lower band signal $\hat{s}_{LB}^{qmf}(n) = \hat{s}_{LB}^{post}(n)$ reconstructed at 12 kbit/s *without high-pass filtering*.
- Above 14 kbit/s (Layers 1 to 4+): In addition to the narrowband CELP and TDBWE decoding, the TDAC decoder reconstructs MDCT coefficients $\hat{D}_{LB}^w(k)$ and $\hat{S}_{HB}(k)$, which correspond to the reconstructed weighted difference in lower band (0-4000 Hz) and the reconstructed signal in higher band (4000-7000 Hz). Note that in the higher band, the non-received sub-bands and the sub-bands with zero bit allocation in TDAC decoding are replaced by the level-adjusted sub-bands of $\hat{S}_{HB}^{bwe}(k)$. Both $\hat{D}_{LB}^w(k)$ and $\hat{S}_{HB}(k)$ are transformed into time domain by inverse MDCT and overlap-add. The lower-band signal $\hat{d}_{LB}^w(n)$ is then processed by the inverse perceptual weighting filter $W_{LB}(z)^{-1}$. To attenuate transform coding artefacts, pre/post-echoes are detected and reduced in both the lower- and higher-band signals $\hat{d}_{LB}(n)$ and $\hat{s}_{HB}(n)$. The lower-band synthesis $\hat{s}_{LB}(n)$ is postfiltered, while the higher-band synthesis $\hat{s}_{HB}^{fold}(n)$ is spectrally folded by $(-1)^n$. The signals $\hat{s}_{LB}^{qmf}(n) = \hat{s}_{LB}^{post}(n)$ and $\hat{s}_{HB}^{qmf}(n)$ are then combined and upsampled in the QMF synthesis filterbank.

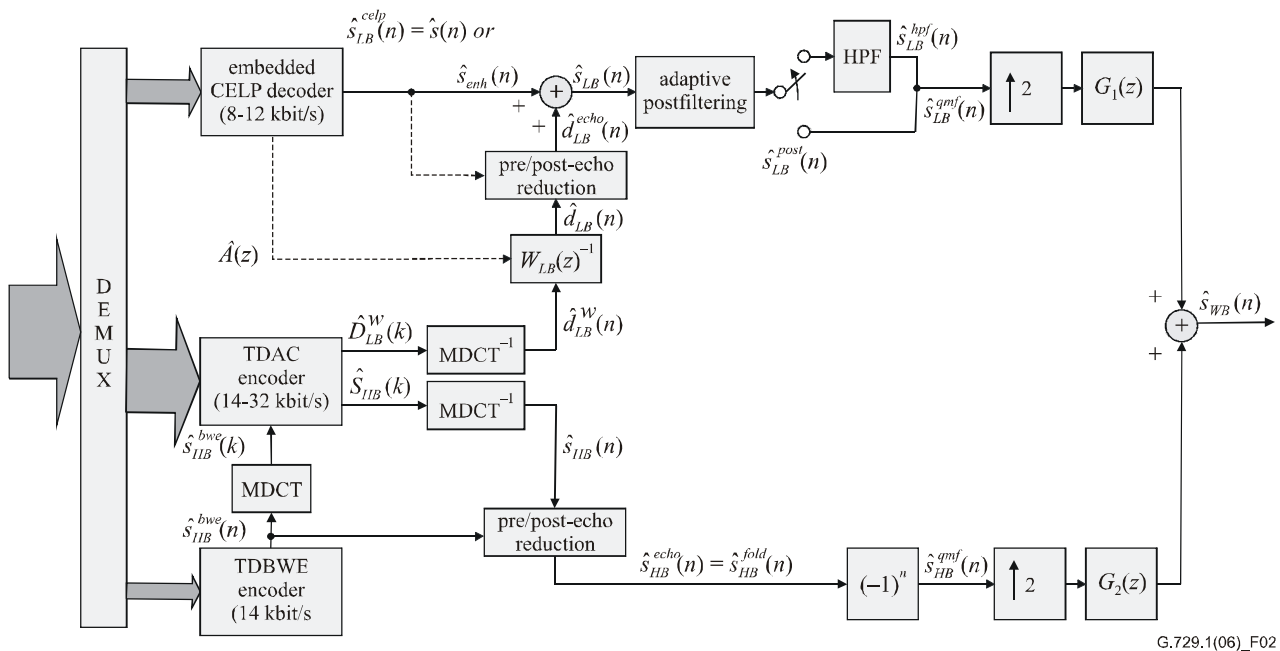


Figure 2 – High-level block diagram of the decoder

5.3 Frame erasure concealment (FEC)

In order to improve the frame erasure concealment as well as convergence and recovery after erased frames, few supplementary concealment/recovery parameters are determined and transmitted. With this encoder-side redundancy, the frame erasure concealment and recovery at the decoder can be significantly improved, especially by improving the convergence of the decoded signal to the actual signal and alleviating the effect of mismatch between the encoder and decoder when normal processing recovers.

The concealment/recovery parameters include signal classification information, energy information, and phase information (the estimated position of the last glottal pulse in the previous superframe). These FEC parameters are estimated using only the lower-band input signal. They are transmitted by the encoder. The classification information requires 2 bits which are transmitted in Layer 2. The phase information is quantized using 7 bits and transmitted in Layer 3. The energy information is quantized using 5 bits and transmitted in Layer 4. The computation of classification, phase, and energy information is described in clauses 6.3.11, 6.5.4 and 6.6.1, respectively.

Note that the class information is not available in case only Layer 1 (core layer) is received. In this case, the signal classification is estimated at the decoder (see 7.6.2).

Frame erasure concealment is performed at the decoder by making use of the transmitted concealment/recovery parameters and by exploiting the extra superframe delay at the decoder. Efficient concealment and recovery techniques are used including glottal pulse resynchronization, energy control and artificial onset reconstruction. This is detailed in 7.6.

5.4 Coder modes

The G.729EV coder is based on a split-band coding approach which naturally yields a very flexible architecture. This coder can easily deal with input and output signals sampled not only at 16000 Hz, but also at 8000 Hz by taking advantage of the QMF analysis and synthesis filterbanks. Table 3 lists the available modes in G.729EV. The DEFAULT mode corresponds to the default operation mode of G.729EV, in which case input and output signals are sampled at 16000 Hz.

Two additional encoder modes are provided:

- The NB_INPUT mode specifies that the encoder input is sampled at 8000 Hz, which allows the bypassing of the QMF analysis filterbank.
- In G729_BST mode, the encoder runs at 8 kbit/s and generates a bitstream with G.729 format using 10 ms frames. The encoder input is sampled at 16000 Hz by default; if the NB_INPUT mode is also set, this input is sampled at 8000 Hz.

On the other hand, three decoder modes are also available:

- The NB_OUTPUT mode specifies that the decoder output is sampled at 8000 Hz which allows the bypassing of the QMF synthesis filterbank.
- In G729B_BST mode the decoder can read and decode G729B frames.
- The LOW_DELAY mode is provided for narrowband use cases. In this case, the decoder bit rate can be limited to 8-14 kbit/s which allows the reduction of the overall algorithmic delay by skipping the inverse MDCT and overlap-add.

In G729B_BST or LOW_DELAY modes, the decoder output is sampled at 16000 Hz by default; if the NB_OUTPUT mode is also set, this output is sampled at 8000 Hz.

Note that wideband quality of the low-delay mode is limited, because scalability cannot be used to increase the bit rate beyond 14 kbit/s. Furthermore, in error-free conditions, the low-delay mode at 14 kbit/s may imply slight quality degradation compared with the default G.729.1 mode at 14 kbit/s.

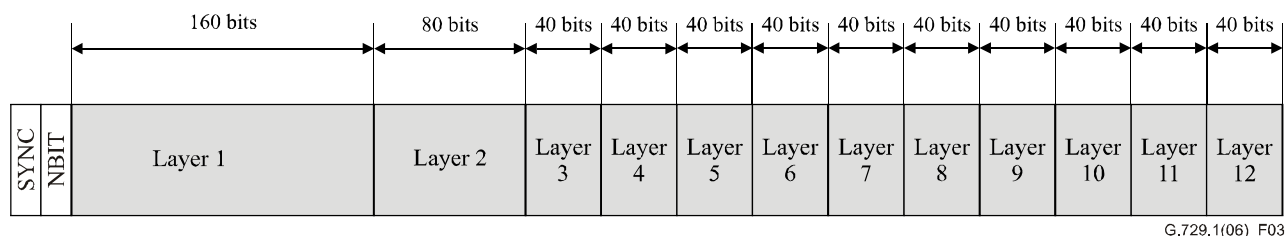
Table 3 – Encoder/decoder modes

Mode	Encoder operation	Decoder operation
DEFAULT	16000 Hz input	16000 Hz output
NB_INPUT	8000 Hz input	N/A
G729_BST	bit rate limited to 8 kbit/s, output G.729 bitstream	N/A
NB_OUTPUT	N/A	8000 Hz output
G729B_BST	N/A	read and decode G729B bitstream
LOW_DELAY	N/A	bit rate limited to 8-14 kbit/s, low delay

5.5 Bit allocation to coder parameters and bitstream layer format

The bit allocation of the coder is presented in Table 4. This table is structured according to the different layers. For a given bit rate, the bitstream is obtained by concatenation of the contributing layers. For example, at 24 kbit/s, which corresponds to 480 bits per superframe, the bitstream comprises Layer 1 (160 bits) + Layer 2 (80 bits) + Layer 3 (40 bits) + Layers 4 to 8 (200 bits).

The G.729EV bitstream format is depicted in Figure 3.



G.729.1(06)_F03

Figure 3 – G.729EV bitstream format (compliant with G.192)

Since the TDAC coder employs spectral envelope entropy coding and adaptive sub-band bit allocation, the TDAC parameters are encoded with a variable number of bits. However, the bitstream above 14 kbit/s can be still formatted into layers of 2 kbit/s, because the TDAC encoder always performs a bit allocation on the basis of the maximum encoder bitrate (32 kbit/s) and the TDAC decoder can handle bitstream truncations at arbitrary positions.

Table 4 – Bit allocation (per 20 ms superframe)

Parameter	Codeword	Number of bits				Total per super-frame
Layer 1 – core layer (narrowband embedded CELP)						
		10 ms frame 1		10 ms frame 2		
Line spectrum pairs	$L0, L1, L2, L3$	18		18		36
		subframe 1	subframe 2	subframe 1	subframe 2	
Adaptive-codebook delay	$P1, P2$	8	5	8	5	26
Pitch-delay parity	$P0$	1		1		2
Fixed-codebook index	$C1, C2$	13	13	13	13	52
Fixed-codebook sign	$S1, S2$	4	4	4	4	16
Codebook gains (stage 1)	$GA1, GA2$	3	3	3	3	12
Codebook gains (stage 2)	$GB1, GB2$	4	4	4	4	16
8 kbit/s core total						160
Layer 2 – narrowband enhancement layer (embedded CELP)						
2nd Fixed-codebook index	$C'1, C'2$	13	13	13	13	52
2nd Fixed-codebook sign	$S'1, S'2$	4	4	4	4	16
2nd Fixed-codebook gain	$G'1, G'2$	3	2	3	2	10
FEC bits (class information)	$CL1, CL2$		1		1	2
12 kbit/s layer total						80
Layer 3 – wideband enhancement layer (TDBWE)						
Time envelope mean	MU	5				5
Time envelope split VQ	$T1, T2$	7+7				14
Frequency envelope split VQ	$F1, F2, F3$	5+5+4				14
FEC bits (phase information)	PH	7				7
14 kbit/s layer total						40
Layers 4 to 12 – wideband enhancement layers (TDAC)						
FEC bits (energy information)	E	5				5
MDCT norm	N	4				4
HB spectral envelope	$RMS2$	variable number $nbits_HB$				$nbits_HB$
LB spectral envelope	$RMS1$	variable number $nbits_LB$				$nbits_LB$
fine structure (VQ of sub-bands coefficients)	$VQ1$ to $VQ18$	$nbits_VQ = 351 - nbits_HB - nbits_LB$				$nbits_VQ$
16-32 kbit/s layer total						360
TOTAL						640

5.6 Algorithmic delay

In default operation (DEFAULT mode), the G.729EV coder has an algorithmic delay of 48.9375 ms (783 samples at 16000 Hz). The delay contributions are listed below:

- 20 ms for input superframe;
- 20 ms for the MDCT analysis (lookahead);
- 5 ms for the narrowband LPC analysis (lookahead);
- 3.9375 ms for the QMF analysis-synthesis filterbank.

Note that for an encoder in NB_INPUT mode and a decoder in NB_OUTPUT and LOW_DELAY mode, the algorithmic delay is reduced to 25 ms. For an encoder in DEFAULT_MODE and a decoder in LOW_DELAY mode at 14 kbit/s, the algorithmic delay is reduced to 28.9375 ms.

5.7 Computational complexity and storage requirement

The observed worst-case complexity of the G.729EV coder (encoder+decoder) is 35.79 WMOPS based on the basic operators of ITU-T Software Tool Library STL2005 v2.1 (from G.191). The G.729EV storage requirement in 16-bit kwords is given in Table 5.

Table 5 – Complexity figures of the G.729EV coder (encoder/decoder)

Computational complexity (WMOPS)	35.7915
Static RAM (kwords)	4.42
Scratch RAM (kwords)	3.74.6
Data ROM (kwords)	8.3
Program ROM (kwords basic operations plus function calls)	328325

5.8 Coder description

The description of the coding algorithm of this Recommendation is made in terms of bit-exact, fixed-point mathematical operations. The ANSI C code indicated in clause 9, which constitutes an integral part of this Recommendation, reflects this bit-exact, fixed-point descriptive approach. The mathematical descriptions of the encoder (clause 6), and decoder (clause 7), can be implemented in several other fashions, possibly leading to a codec implementation not complying with this Recommendation. Therefore, the algorithm description of the ANSI C code of clause 9 shall take precedence over the mathematical descriptions of clauses 6 and 7 whenever discrepancies are found. A non-exhaustive set of test signals, which can be used with ANSI C code, are available from TSB.

6 Functional description of the encoder

6.1 QMF analysis filterbank, signal downscaling and spectral folding of the higher band

A QMF analysis filterbank is used to split the input signal $s_{WB}(n)$ sampled at 16000 Hz into two signals sampled at 8000 Hz: a lower-band signal $s_{LB}^{qmf}(n)$ and a higher-band signal $s_{HB}^{qmf}(n)$, as shown in Figure 1. In the used QMF filterbank, the low-pass filter $H_1(z)$ is a symmetric FIR filter with 64 coefficients. The coefficients of the high-pass filter $H_2(z)$ are related to the coefficients of $H_1(z)$ by:

$$h_2(n) = (-1)^n h_1(n) \quad (1)$$

The frequency responses of the filters $H_1(z)$ and $H_2(z)$ are shown in Figure 4.

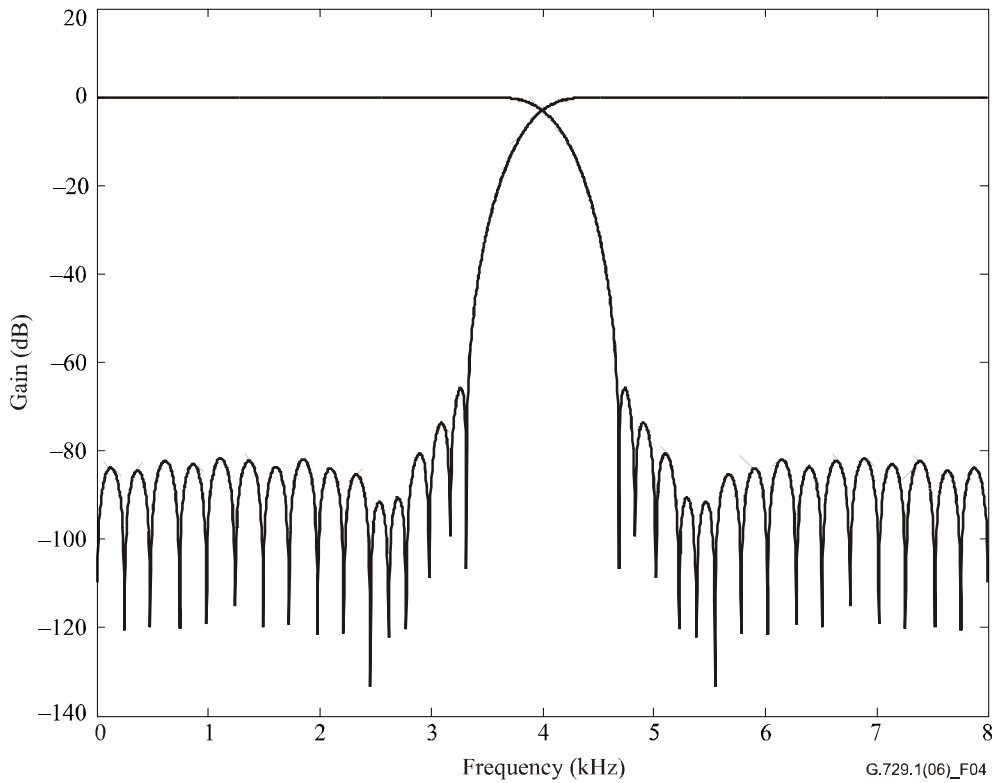


Figure 4 – Frequency responses of $H_1(z)$ and $H_2(z)$

Since the filter is symmetric, only half the coefficients are stored. Let $h_1(j)$, $j = 0, \dots, 31$, denote the second half of the filter coefficients (the whole filter extended from $j = -32$ to $j = 31$ with $h_1(j) = h_1(-j-1)$, $j = 0, \dots, 31$). The lower-band signal sampled at 16000 Hz is given by:

$$y_1(n) = \sum_{j=0}^{31} h_1(j) [s_{WB}(n+1+j) + s_{WB}(n-j)] \quad (2)$$

Similar to Equation 2, the higher-band signal sampled at 16000 Hz is given by:

$$y_2(n) = \sum_{j=0}^{31} h_2(j) [s_{WB}(n+1+j) + s_{WB}(n-j)] \quad (3)$$

Given the relation in Equation 1, and the fact that $h_2(j)$ is anti-symmetric ($h_2(j) = -h_2(-j-1)$), the higher-band signal in Equation 3 can be written as:

$$y_2(n) = \sum_{j=0}^{31} h_1(j) (-1)^j [s_{WB}(n+1+j) + s_{WB}(n-j)] \quad (4)$$

The lower- and higher-band signals sampled at 8000 Hz are simply obtained by dropping one out of two samples. That is:

$$s_{LB}^{qmf}(n/2) = y_1(n) \text{ and } s_{HB}^{qmf}(n/2) = y_2(n).$$

Thus in Equations 2 and 4, the signals $y_1(n)$ and $y_2(n)$ are computed only at values $n = 0, 2, 4, \dots$ (polyphase implementation) to directly obtain $s_{LB}^{qmf}(n/2)$ and $s_{HB}^{qmf}(n/2)$.

Note that given the similarity between Equations 2 and 4, the two signals $y_1(n)$ and $y_2(n)$ can be computed together in order to save complexity. This is done here by splitting the computation according to even and odd coefficient indices. First, the intermediate signals $A(n)$ and $B(n)$ are computed as:

$$A(n) = \sum_{j=0}^{15} (h_1(2j)s_{WB}(n+1+2j) + h_1(2j+1)s_{WB}(n-2j-1)) \quad (5)$$

$$B(n) = \sum_{j=0}^{15} (h_1(2j)s_{WB}(n-2j) + h_1(2j+1)s_{WB}(n+1+2j+1)) \quad (6)$$

and then the signals $y_1(n)$ and $y_2(n)$ are computed as:

$$y_1(n) = A(n) + B(n)$$

$$y_2(n) = A(n) - B(n)$$

Note that the QMF analysis filterbank includes a signal downscaling (by a factor of 2).

The higher-band signal $s_{HB}^{qmf}(n)$ is spectrally folded as follows:

$$s_{HB}^{fold}(n) = (-1)^n s_{HB}^{qmf}(n), \quad n = 0, \dots, 159$$

6.2 Pre-processing of lower band

The lower-band signal is filtered by an elliptic high-pass filter $H_{h1}(z)$ of order 2 to remove the frequency components below 50 Hz. The filter $H_{h1}(z)$ is defined as:

$$H_{h1}(z) = \frac{0.95551031152729 - 1.91102039813878 z^{-1} + 0.95551031152729 z^{-2}}{1 - 1.96646455789804 z^{-1} + 0.9671820760729101 z^{-2}}$$

6.3 Narrowband embedded CELP encoder (Layers 1 and 2)

6.3.1 Linear prediction analysis and quantization

Same as 3.2/G.729.

Note that the interpolated and quantized linear predictive filter $\hat{A}(z)$ is also needed in the TDAC encoder. The results of the LP analysis, quantization and interpolation performed by the CELP encoder are delivered to the TDAC encoder.

6.3.2 Perceptual weighting

Same as 3.3/G.729.

6.3.3 Open-loop pitch analysis

A smoothed open-loop pitch track can help stabilizing the perceptual speech quality. Especially, smoothed pitch track can make pitch prediction (pitch estimation for lost frames) easier when applying frame erasure concealment algorithm at the decoder side. The G.729 open-loop pitch analysis algorithm exploits only current frame information to avoid pitch lag multiples. The smoothing algorithm as described here improves this algorithm by exploiting the open-loop pitch of the previous frame.

The G.729 open-loop pitch analysis algorithm is modified as follows. Every 10 ms frame, the initial open-loop pitch analysis gives 3 candidates from 3 search ranges:

$$\{t_1, R'(t_1)\}, \{t_2, R'(t_2)\}, \{t_3, R'(t_3)\}$$

where $t_1 > t_2 > t_3$ are open-loop pitch candidates and $R'(t_1)$, $R'(t_2)$ and $R'(t_3)$ are the corresponding long-term correlation values. See 3.4/G.729 for the definition of $R'(t_i)$. If $T_{op,old}$ is noted as previous open-loop pitch and δ is a threshold value which will influence the decision to pick up one of the 3 candidates, the pitch candidate close to the previous pitch $T_{op,old}$ is favoured by the following procedure in three steps:

Step 1: Select initial candidate by looking for maximum pitch correlation

$$R'_{\max} = \max_{i=1,2,3} R'(t_i)$$

Let T_{op} be the initial open-loop pitch candidate which corresponds to the above maximum correlation value.

Step 2: if $t_2 < T_{op}$

$$\text{if } |T_{op,old} - t_2| < 10, \delta = 0.7 \text{ else } \delta = 0.9$$

$$\text{if } R'_{\max} \delta < R'(t_2), R'_{\max} = R'(t_2) \text{ and } T_{op} = t_2$$

Step 3: if $t_3 < T_{op}$

$$\text{if } |T_{op,old} - t_3| < 5, \delta = 0.7 \text{ else } \delta = 0.9$$

$$\text{if } R'_{\max} \delta < R'(t_3), T_{op} = t_3$$

The final selection of open-loop pitch for current frame is T_{op} .

6.3.4 Computation of the impulse response

Same as 3.5/G.729.

6.3.5 Computation of the target signal

Same as 3.6/G.729.

6.3.6 Adaptive codebook search

Same as 3.7/G.729.

6.3.7 Fixed codebook at 8 kbit/s – structure and search

The fixed codebook structure is the same as 3.8/G.729. Its search procedure is also similar to the G.729 fixed-codebook search except for some parts, which are described in this clause.

Orthogonalization of the fixed codebook search

The fixed codebook search is performed using the orthogonal search procedure in which filtered fixed-codebook vectors are orthogonalized to the filtered adaptive-codebook vector.

To realize the orthogonal search, Equations 50/G.729 and 51/G.729 are replaced with the following ones. Equation 50/G.729 is redefined by:

$$x'(n) = \left(\sum_{i=0}^{39} y(i)y(i) \right) x(n) - \left(\sum_{i=0}^{39} x(i)y(i) \right) y(n) \quad n = 0, \dots, 39 \quad (7)$$

and Equation 51/G.729 is changed to:

$$\phi(i, j) = \left(\sum_{n=0}^{39} y(n)y(n) \right) \left(\sum_{n=j}^{39} h(n-i)h(n-j) \right) - y'(i)y'(j) \quad i = 0, \dots, 39 \quad j = i, \dots, 39 \quad (8)$$

The correlation signal $y'(n)$ is obtained from the filtered adaptive-codebook vector $y(n)$ and the impulse response $h(n)$ by:

$$y'(39-i) = \sum_{j=0}^i h(j)y(39-i+j) \quad i = 0, \dots, 39 \quad (9)$$

Using the above target vector $x'(n)$ and the matrix Φ_{ij} , the orthogonal search is accomplished through the procedure described below.

Fast codebook search using global pulse replacement

The fast search procedure described in this clause only concerns the pulse positions. The pulse amplitudes (signs) are pre-selected using the procedure described in 3.8.1/G.729.

The fixed codebook is searched based on global pulse replacement method. This procedure consists of two stages: initial codevector determination and pulse replacement. The initial codevector is determined by the pulse positions which maximize in each track the pulse-position likelihood-estimate vector $b(n)$ given by:

$$b(n) = \frac{d(n)}{\sqrt{\sum_{i=0}^{39} d(i)d(i)}} + \frac{r_{LTP}(n)}{\sqrt{\sum_{i=0}^{39} r_{LTP}(i)r_{LTP}(i)}} \quad n = 0, \dots, 39 \quad (10)$$

where $r_{LTP}(n)$ is the long-term prediction residual signal given by:

$$r_{LTP}(n) = r(n) - g_p v(n) \quad n = 0, \dots, 39 \quad (11)$$

where $r(n)$ is the LP residual signal of the input $s(n)$, g_p is the pitch gain computed as explained in Equation 43/G.729 and $v(n)$ is the adaptive codebook vector computed in Equation 40/G.729.

After the initial codevector determination, the optimum codevector is searched by replacing the initial pulse of each track with a new pulse which maximizes C_k^2/E_k of Equation 53/G.729. The pulse replacement procedure is repeated 4 times or terminated when the C_k^2/E_k does not increase. For example, assume that the initial codevector is determined as $(m_{01}, m_{11}, m_{21}, m_{31})$. In the symbol m_{xy} , x is the track number and y is the pulse position in each track. At the first step, a new codevector is searched by maximizing the C_k^2/E_k on the following codevectors.

Track 0: $(m_{02}, m_{11}, m_{21}, m_{31}), (m_{03}, m_{11}, m_{21}, m_{31}), \dots, (m_{08}, m_{11}, m_{21}, m_{31})$

Track 1: $(m_{01}, m_{12}, m_{21}, m_{31}), (m_{01}, m_{13}, m_{21}, m_{31}), \dots, (m_{01}, m_{18}, m_{21}, m_{31})$

Track 2: $(m_{01}, m_{11}, m_{22}, m_{31}), (m_{01}, m_{11}, m_{23}, m_{31}), \dots, (m_{01}, m_{11}, m_{28}, m_{31})$

Track 3: $(m_{01}, m_{11}, m_{21}, m_{32}), (m_{01}, m_{11}, m_{21}, m_{33}), \dots, (m_{01}, m_{11}, m_{21}, m_{316})$

If the C_k^2/E_k is maximized at codevector $(m_{03}, m_{11}, m_{21}, m_{31})$ and greater than that of initial codevector $(m_{01}, m_{11}, m_{21}, m_{31})$, the codevector $(m_{03}, m_{11}, m_{21}, m_{31})$ is selected as a new codevector. As a result of the first step, m_{01} of the initial codevector is replaced with m_{03} . In the second step, the same procedure is repeated on the following codevectors except for track 0.

Track 1: $(m_{03}, m_{12}, m_{21}, m_{31}), (m_{03}, m_{13}, m_{21}, m_{31}), \dots, (m_{03}, m_{18}, m_{21}, m_{31})$

Track 2: $(m_{03}, m_{11}, m_{22}, m_{31}), (m_{03}, m_{11}, m_{23}, m_{31}), \dots, (m_{03}, m_{11}, m_{28}, m_{31})$

Track 3: $(m_{03}, m_{11}, m_{21}, m_{32}), (m_{03}, m_{11}, m_{21}, m_{33}), \dots, (m_{03}, m_{11}, m_{21}, m_{316})$

A new codevector is selected in the same manner of the first step. As the pulse replacement procedure is repeated, the ratio C_k^2/E_k of the replaced codevector is increased. In this new approach a smaller number of pulse position combinations are tested and it has less complexity.

6.3.8 Quantization of the 8 kbit/s gains

Same as 3.9/G.729.

6.3.9 Fixed codebook at 12 kbit/s – structure and search

The 12 kbit/s enhancement layer relies on an extra fixed codebook that, when combined with the 8 kbit/s core layer excitation, offers a richer excitation signal. In particular, this extra fixed codebook better represents high frequencies of the LP excitation. This benefit is increased by the use of a modified perceptual filter to stress high frequencies.

Computation of the 12 kbit/s target signal

The target signal $x_{enh}(n)$ for the 12 kbit/s FCB search is first obtained by filtering the difference between the LP residual $r(n)$ and excitation $u(n)$ of Equation 75/G.729 at the 8 kbit/s layer through the filters $1/\hat{A}(z)$ and $W(z)$:

$$x_{enh}(z) = \frac{W(z)}{\hat{A}(z)}(r(z) - u(z)) \quad (12)$$

Structure of the codebook

The 12 kbit/s fixed codebook $\mathcal{C}'$ is defined by a *tri-pulse pattern* $-\alpha_{enh}z^{-1} + 1 - \alpha_{enh}z$, with a central pulse +1 and two side pulses with lower magnitude and opposite sign $-\alpha_{enh}$.

Each codevector is obtained by adding 4 occurrences of this pattern scaled by a sign factor ± 1 . The centres of the pattern occurrences occupy the same sets of positions as the G.729 fixed ACELP codebook pulses, recalled in Table 6.

Table 6 – Structure of the extra fixed codebook

Pattern occurrence	Sign	Positions
i_0	$s'_0: \pm 1$	m'0: 0, 5, 10, 15, 20, 25, 30, 35
i_1	$s'_1: \pm 1$	m'1: 1, 6, 11, 16, 21, 26, 31, 36
i_2	$s'_2: \pm 1$	m'2: 2, 7, 12, 17, 22, 27, 32, 37
i_3	$s'_3: \pm 1$	m'3: 3, 8, 13, 18, 23, 28, 33, 38 4, 9, 14, 19, 24, 29, 34, 39

The codebook vector $c'(n)$ is constructed by taking a zero vector of dimension 40, and adding the four occurrences of the pattern, truncated at the border if necessary to fit the block size, and multiplied with their corresponding sign:

$$c'(n) = \sum_{i=0}^3 s'_i \times (-\alpha_{enh} \delta(n - m'_i + 1) + \delta(n - m'_i) - \alpha_{enh} \delta(n - m'_i - 1)) \quad n = 0, \dots, 39 \quad (13)$$

where $\delta(n)$ is a unit pulse.

Note that, in a way analogous to the adaptive pre-filtering of the fixed codebook search in the core layer, the tri-pulse pattern can be combined with the impulse response of the perceptual weighting filter to use the ordinary algebraic codebook search procedure, as will be shown below (from Equation 13, the tri-pulse pattern is equivalent to an adaptive pre-filter of the form $P'(z) = -\alpha_{enh}z + 1 - \alpha_{enh}z^{-1}$).

Tri-pulse pattern adaptation

The coefficient α_{enh} is adapted to the nature of the signal. Modification of α_{enh} allows the tri-pulse shape to change. For each 5 ms subframe, α_{enh} is given by:

$$\alpha_{enh} = 0.17(1 + r_v) \quad (14)$$

where r_v is related to the voicing nature of the signal. Here, r_v is given by

$$r_v = (E_v - E_c) / (E_v + E_c) \quad (15)$$

where E_v and E_c are the energies of the scaled pitch codevector $\hat{g}_p v(n)$ and scaled fixed codevector $\hat{g}_c c(n)$, respectively. Note that since the value of r_v is between -1 and 1 , the value of α_{enh} is between 0 and 0.34 . Thus, the factor α_{enh} is related to the amount of voicing with a value of 0 for purely unvoiced segments and a value of 0.34 for purely voiced segments. Therefore, the 12 kbit/s fixed codebook has a higher frequency content in voiced signals and less for unvoiced signals.

Codebook search

The 12 kbit/s fixed codebook is searched by minimizing the mean-square error between the target signal $x_{enh}(n)$ and a codebook excitation filtered through the modified perceptual filter to stress high frequencies. The perceptual filter is modified by applying the high-pass filter $-0.15z^{-1} + 1 - 0.15z$ to $W(z)$.

This modification of the perceptual filter is introduced by modifying the target vector and the weighting filter impulse response.

The modification of the target signal is obtained by:

$$x'_{enh}(n) = -0.15x_{enh}(n-1) + x_{enh}(n) - 0.15x_{enh}(n+1) \quad n = 0, \dots, 39 \quad (16)$$

where $x_{enh}(-1) = x_{enh}(40) = 0$

The impulse response $h_{enh}(n)$ of the modified weighting filter is obtained by:

$$h_{enh}(n) = -0.15h(n-1) + h(n) - 0.15h(n+1) \quad n = 0, \dots, 39 \quad (17)$$

where $h(-1) = h(40) = 0$

where $h(n)$ is the impulse response computed as in 3.5/G.729. The search procedure exploits the fact that the occurrences of the tri-pulse pattern are centered on the same positions as the G.729 fixed codebook pulses. It is possible to transfer the effect of the tri-pulse pattern on the impulse

response $h_{enh}(n)$. Border effects due to the possible truncation of the codevectors at the block boundaries may be neglected, since the magnitude of the side pulses is maintained much lower than the central pulse. With such a modification the 12 kbit/s fixed codebook search becomes quite similar to 8 kbit/s fixed codebook search.

To this end, the impulse response $h_{enh}(n)$ is modified according to:

$$h'_{enh}(n) = -\alpha_{enh} h_{enh}(n-1) + h_{enh}(n) - \alpha_{enh} h_{enh}(n+1) \quad n = -1, \dots, 39 \quad (18)$$

$$\text{where } h_{enh}(-2) = h_{enh}(-1) = h_{enh}(40) = 0$$

The $\mathbf{H}_{enh}$ matrix is defined as a convolution matrix:

$$\mathbf{H}_{enh} = \begin{bmatrix} h'_{enh}(0) & h'_{enh}(-1) & 0 & \dots & 0 \\ h'_{enh}(1) & h'_{enh}(0) & h'_{enh}(-1) & \ddots & \vdots \\ h'_{enh}(2) & h'_{enh}(1) & h'_{enh}(0) & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & h'_{enh}(-1) \\ h'_{enh}(39) & h'_{enh}(38) & h'_{enh}(37) & \dots & h'_{enh}(0) \end{bmatrix} \quad (19)$$

The matrix $\Phi_{enh} = \mathbf{H}_{enh}^t \mathbf{H}_{enh}$ contains the correlations of $h'_{enh}(n)$, and the elements of the matrix $\Phi_{enh} = \{\phi_{enh}(i, j)\}$ are given by:

$$\begin{aligned} \phi_{enh}(0,0) &= \sum_{n=0}^{39} h'_{enh}(n)^2 \\ \phi_{enh}(i,j) &= \phi_{enh}(j,i) = \sum_{n=-1}^{39-j} h'_{enh}(n+j-i) h'_{enh}(n) \quad i=0, \dots, 39; \quad j=i, \dots, 39 \end{aligned} \quad (20)$$

The correlation signal $d_{enh}(n)$ is obtained from the target signal $x_{enh}(n)$ and the modified impulse response $h'_{enh}(n)$ by:

$$\begin{aligned} d_{enh}(0) &= \sum_{n=0}^{39} x'_{enh}(n) h'_{enh}(n) \\ d_{enh}(i) &= \sum_{n=-1}^{39-i} x'_{enh}(n+i) h'_{enh}(n) \quad i=1, \dots, 39 \end{aligned} \quad (21)$$

The codebook is searched by maximizing the term

$$\frac{C'^2}{E'} = \frac{\left(\sum_{i=0}^3 s'_i \times d_{enh}(m'_i) \right)^2}{\sum_{i=0}^3 \phi_{enh}(m'_i, m'_i) + 2 \sum_{i=0}^2 \sum_{j=i+1}^3 s'_i \times s'_j \times \phi_{enh}(m'_i, m'_j)} \quad (22)$$

among the codevectors defined by the positions $\{m'_i, i=0, \dots, 3\}$ and signs $\{s'_i, i=0, \dots, 3\}$ of their 4-pattern occurrences.

The 12 kbit/s fixed codebook search is then performed in the same manner as the 8 kbit/s fixed-codebook search.

Index computation

The codevector indices and signs are encoded as the 8 kbit/s fixed codebook indices and signs (defined in 3.8.2/G.729) with the signs s'_i and the positions of centres m'_i of the pattern occurrences replacing the signs s_i and positions m_i of the 8 kbit/s fixed codebook pulses.

6.3.10 Quantization of the 12 kbit/s fixed codebook gain

The quantization of the 12 kbit/s fixed codebook gain is performed by minimizing the mean-square weighted error between the target signal $x_{enh}(n)$ and the filtered codebook excitation $z_{enh}(n)$, which is the fixed-codebook vector $c'(n)$ convolved with $h_{enh}(n)$:

$$z_{enh}(n) = \sum_{k=0}^{39} c'(k) \times h_{enh}(n-k) \quad (23)$$

With (13) this is equivalent to

$$z_{enh}(n) = \sum_{i=0}^3 s'_i \times h'_{enh}(n-m'_i) \quad (24)$$

As in ITU-T Rec. G.729, this vector is obtained at the end of the codebook search. First the cross-correlation and energy terms xy and yy are calculated:

$$xy = \sum_{i=0}^{39} z_{enh}(n) \times x'_{enh}(n) \quad (25)$$

$$yy = \sum_{i=0}^{39} z_{enh}(n)^2 \quad (26)$$

The 12 kbit/s fixed-codebook gain g_{enh} is scalar quantized relatively to the 8 kbit/s quantized fixed codebook gain $\hat{g}_c$. The number of bits depends on the subframe index:

- For the first subframe of each 10 ms frame, the ratio $g_{enh} / \hat{g}_c$ is scalar quantized into $\hat{g}_{enh} / \hat{g}_c$ with 3 bits using a Lloyd-Max quantizer and the weighted mean-square criterion. The optimal ratio $\hat{g}_{enh} / \hat{g}_c$ minimizes:

$$-2 \, xy \times \frac{\hat{g}_{enh}}{\hat{g}_c} \times \hat{g}_c + yy \times \left(\frac{\hat{g}_{enh}}{\hat{g}_c} \right)^2 \times (\hat{g}_c)^2 \quad (27)$$

- For the second subframe of each 10 ms frame, the ratio $g_{enh} / \hat{g}_c$ is uniformly quantized in dB with 2 bits in the [−6 dB, 3 dB] range using the mean-square error criterion.

6.3.11 Signal classification for FEC

The basic idea behind using a classification of the speech for a signal reconstruction in the presence of erased superframes consists in the fact that the ideal concealment strategy is different for quasi-stationary speech segments and for speech segments with rapidly changing characteristics. Whereas the best processing of erased superframes in non-stationary speech segments can be summarized as a rapid drop of energy, in the case of quasi-stationary signal, the speech-encoding parameters do not vary dramatically and can be kept practically unchanged during several adjacent erased superframes before being damped. Thus, the optimal method for a signal recovery following an erased block of superframes varies with the classification of the speech signal.

Note that this information is not available if only the core 8 kbit/s layer is received. In this case, signal classification is performed at the decoder as will be explained in 7.6.1.

The speech signal can be roughly classified as voiced, unvoiced and pauses. Voiced speech contains an important amount of periodic components and can be further divided into the following categories: voiced onsets, voiced segments, voiced transitions and voiced offsets. A voiced onset is defined as a beginning of a voiced speech segment after a pause or an unvoiced segment. During voiced segments, the speech signal parameters (spectral envelope, pitch period, ratio of periodic and non-periodic components, energy) vary slowly from superframe to superframe. A voiced transition is characterized by rapid variations of a voiced speech, such as a transition between vowels. Voiced offsets are characterized by a gradual decrease of energy and voicing at the end of voiced segments.

The unvoiced parts of the signal have no periodic components and can be further divided into unstable superframes, where the energy and the spectrum change rapidly, and stable superframes where these characteristics remain relatively stable. Remaining superframes are classified as silence. Silence superframes comprise all superframes without active speech, i.e., also noise-only superframes if a background noise is present.

The above-mentioned classes do not need a separate concealment. Some of the signal classes are grouped together.

Signal classes

The superframe classification is done with the consideration of the concealment and recovery strategy in mind. In other words, any superframe is classified in such a way that the concealment can be optimal if the following superframe is missing, or that the recovery can be optimal if the previous superframe was lost. Some of the classes used for the FEC processing does not need to be transmitted, as they can be deduced without ambiguity at the decoder. Here, five distinct classes are used and defined as follows:

- UNVOICED class comprises all unvoiced speech superframes and all superframes without active speech. A voiced offset superframe can be also classified as UNVOICED if its end tends to be unvoiced and the concealment designed for unvoiced superframes can be used for the following superframe in case it is lost.
- UNVOICED TRANSITION class comprises unvoiced superframes with a possible voiced onset at the end. The onset is however still too short, or not built well enough to use the concealment designed for voiced superframes. The UNVOICED TRANSITION class can only follow a superframe classified as UNVOICED or UNVOICED TRANSITION.
- VOICED TRANSITION class comprises voiced superframes with relatively weak-voiced characteristics. Those are typically voiced superframes with rapidly changing characteristics (transitions between vowels) or voiced offsets lasting the whole superframe. The VOICED TRANSITION class can only follow a superframe classified as VOICED TRANSITION, VOICED or ONSET.
- VOICED class comprises voiced superframes with stable characteristics. This class can only follow a superframe classified as VOICED TRANSITION, VOICED or ONSET.
- ONSET class comprises all voiced superframes with stable characteristics following a superframe classified as UNVOICED or UNVOICED TRANSITION. Superframes classified as ONSET correspond to voiced onset superframes where the onset is already sufficiently well built for the use of the concealment designed for lost voiced superframes. The concealment techniques used for a frame erasure following the ONSET class are the same as following the VOICED class. The difference is in the recovery strategy. If an ONSET class superframe is lost (i.e., a VOICED good superframe arrives after an erasure, but the last good superframe before the erasure was UNVOICED), a special technique can be used to artificially reconstruct the lost onset. The artificial onset reconstruction techniques will be described in 7.6.7. On the other hand, if an ONSET good superframe arrives after an erasure and the last good superframe before the erasure was UNVOICED,

this special processing is not needed, as the onset has not been lost (has not been in the lost superframe).

It should be noted that UNVOICED TRANSITION class and VOICED TRANSITION class can be grouped together as they can be unambiguously differentiated at the decoder (UNVOICED TRANSITION can only follow UNVOICED or UNVOICED TRANSITION superframes, VOICED TRANSITION can only follow ONSET, VOICED or VOICED TRANSITION superframes). Thus, although there are five classes, only 2 bits are needed.

Class information

The 2 bits of class information describe the following 4 classes: VOICED, UNVOICED, ONSET, TRANSITION. At the decoder, if a TRANSITION class is received, it is interpreted as UNVOICED TRANSITION if the class of the preceding superframe is UNVOICED or UNVOICED TRANSITION. Otherwise it is interpreted as VOICED TRANSITION.

Signal classification parameters

The following parameters are used for the classification at the encoder: a normalized correlation $\bar{r}_{sw}$, a spectral tilt measure e_t , a signal-to-noise ratio snr , a pitch stability counter pc , a relative superframe energy of the signal at the end of the current superframe E_s , and a zero-crossing counter zc . The computation of these parameters which are used to classify the signal is explained below.

Normalized correlation

The normalized correlation r_x is computed as part of the open-loop pitch search which outputs the open-loop pitch estimate every 10 ms (twice per superframe). Here, it is also used to output the normalized correlation measures. These normalized correlations are computed on the current weighted speech signal $s_w(n)$ at the open-loop pitch delay. The average correlation $\bar{r}_{sw}$ is defined as

$$\bar{r}_{sw} = 0.5(r_{sw}(1) + r_{sw}(2)) \quad (28)$$

where $r_{sw}(1)$, $r_{sw}(2)$ are respectively the normalized correlation of the first 10 ms frame and second 10 ms frame. The normalized correlation $r_{sw}(i)$ is computed as follows:

$$r_{sw}(i) = \frac{\sum_{n=0}^{79} s_w(80(i-1) + n) s_w(80(i-1) + n - T_{op}(i))}{\sqrt{\sum_{n=0}^{79} s_w^2(80(i-1) + n) \sum_{n=0}^{79} s_w^2(80(i-1) + n - T_{op}(i))}} \quad (29)$$

The correlations $r_{sw}(i)$ are computed using the weighted speech signal $s_w(n)$. $T_{op}(i)$ is the open-loop pitch computed in the i -th 10 ms frame as described in 6.3.3.

Spectral tilt

The spectral tilt is estimated as the normalized first autocorrelation coefficient of the speech signal (the first reflection coefficient obtained during LP analysis). Since LP analysis is performed twice per superframe (once every 10 ms frame), the spectral tilt is computed as the average of the first reflection coefficient from both LP analysis. That is

$$e_t = -0.5(k_1(1) + k_1(2)) \quad (30)$$

where $k_1(i)$ is the first reflection coefficient from the LP analyses in the i -th frame computed as in 3.2.2/G.729.

Signal-to-noise ratio

The signal to noise ratio (SNR) measure exploits the fact that for a general waveform matching encoder, the SNR is much higher for voiced sounds. The *snr* parameter is computed for the whole superframe using the relation:

$$snr = \frac{E_s}{E_d} \quad (31)$$

where E_s is the energy of the speech signal $s(n)$ in the current superframe and E_d is the energy of the difference $d_{LB}(n)$ between the speech signal $s(n)$ and the local synthesis signal $\hat{s}_{enh}(n)$ at 12 kbit/s in the current superframe.

Pitch stability counter

The pitch stability counter *pc* assesses the variation of the pitch period. It is computed as follows:

$$pc = |T_2(2) - T_1(2)| + |T_1(2) - T_2(1)| \quad (32)$$

where the values $T_1(i)$ and $T_2(i)$ correspond respectively to the first and second closed-loop pitch lags of the *i*-th 10 ms frame, i.e., *pc* is computed with the closed-loop pitch lags of the last three 5 ms subframes.

Relative energy

The relative superframe energy E_s is computed as a difference between the superframe energy in dB and its long-term average

$$E_s = E_t - E_{lt} \quad (33)$$

where the superframe energy E_t is the energy of the windowed input signal in dB

$$E_t = 10 \log_{10} \left(\frac{1}{160} \sum_{n=0}^{159} [s(n)w_{Hanning}(n)]^2 \right) \quad (34)$$

where $w_{Hanning}(n)$ is a Hanning window of length 160:

$$w_{Hanning}(n) = \frac{1}{2} \left(1 - \cos \left(\frac{2\pi n}{159} \right) \right), \quad n = 0, \dots, 159 \quad (35)$$

The long-term averaged energy is updated during active speech superframes using the following relation:

$$E_{lt} = 0.99E_{lt} + 0.01E_t \quad (36)$$

Zero crossing

The last parameter is the zero-crossing parameter *zc* computed on the current speech superframe. Here, the zero-crossing counter *zc* counts the number of times the signal sign changes from positive to negative during that interval.

...

Classification procedure

The classification parameters are used to define a function of merit f_m . For that purpose, the classification parameters are first scaled between 0 and 1 so that each parameter translates in 0 for an unvoiced signal and 1 for a voiced signal. Each parameter p_x is scaled by a linear function as follows:

$$p_x^s = k_x p_x + c_x \quad (37)$$

and clipped between 0 and 1 (except for the relative energy which is clipped between 0.5 and 1). The function coefficients k_x and c_x have been found experimentally for each of the parameters so that the signal distortion due to the concealment and recovery techniques used in presence of superframe erasures is minimal. The values used are summarized in Table 7.

Table 7 – Scaling function coefficients of the classification parameters

Parameter	Meaning	k_x	c_x
$\bar{r}_{sw}$	Normalized correlation	2.857	-1.286
e_t	Spectral tilt	2.5	-1.25
snr	Signal to noise ratio	0.09615	-0.25
pc	Pitch stability counter	-0.1176f	2.0
E_s	Relative frame energy	0.05	0.45
zc	Zero crossing counter	-0.1	2.0

The function of merit has been defined as:

$$f_m = \frac{1}{7} (2\bar{r}_{sw}^s + e_t^s + 1.2snr^s + pc^s + E_s^s + zc^s) \quad (38)$$

where the superscript s indicates the scaled version of the parameters.

The function of merit is then scaled by 1.05 if $E_s^s \leq 0.5$ and scaled by 1.25 if $E_s^s > 0.75$. Further, the function of merit is also scaled by a factor f_E based on a state machine which checks the difference between the instantaneous relative energy variation and the long-term relative energy variation. This is added to improve the signal classification in the presence of background noise. To do so, a relative energy variation parameter E_{var} is updated as:

$$E_{var} = 0.05(E_s - E_{prev}) + 0.95E_{var} \quad (39)$$

where E_{prev} is the value of E_s from previous superframe.

if $|E_s - E_{prev}| < (|E_{var}| + 6)$ and ($class_{old} = \text{UNVOICED}$) $f_E = 0.8$

else

if $(E_s - E_{prev}) > (E_{var} + 3)$ and ($class_{old} = \text{UNVOICED}$ or TRANSITION) $f_E = 1.1$

else

if $(E_s - E_{prev}) < (E_{var} - 5)$ and ($class_{old} = \text{VOICED}$ or ONSET) $f_E = 0.6$.

where $class_{old}$ is the class of the previous superframe.

The classification is then done using the function of merit f_m and following the rules summarized in Table 8.

Table 8 – Signal classification rules at the encoder

Previous frame class	Rule	Current superframe class
ONSET VOICED VOICED TRANSITION	$f_m \geq 0.68$	VOICED
	$0.56 \leq f_m < 0.68$	VOICED TRANSITION
	$f_m < 0.56$	UNVOICED
UNVOICED TRANSITION UNVOICED	$f_m > 0.64$	ONSET
	$0.64 \geq f_m > 0.58$	UNVOICED TRANSITION
	$f_m \leq 0.58$	UNVOICED

For the purpose of FEC classification, inactive speech superframes are classified as UNVOICED. A superframe is directly classified as UNVOICED if the relative energy is less than -10 dB (as a simple measure for speech inactivity).

6.3.12 Memory update

Same as 3.10/G.729

In addition, the states of the synthesis and weighting filters for the 12 kbit/s target signal in the next subframe are updated. This memory update is the same as in G.729 except that it is based on the excitation signal reconstructed at 12 kbit/s instead of excitation signal reconstructed at 8 kbit/s.

6.4 Pre-processing of the higher band

The higher-band signal is filtered by an elliptic low-pass filter $H_{h2}(z)$ of order 4 to remove the frequency components in the 3000-4000 Hz band (which corresponds to the band 7000-8000 Hz in the original signal). The filter $H_{h2}(z)$ is defined as:

$$H_{h2}(z) = \frac{0.3500277721 + 1.3045646694z^{-1} + 1.9127698530z^{-2} + 1.3045646694z^{-3} + 0.3500277721z^{-4}}{1 + 1.79857371201z^{-1} + 1.69962113314z^{-2} + 0.70669663302z^{-3} + 0.16954708937z^{-4}} \quad (40)$$

6.5 TDBWE encoder (Layer 3)

The TDBWE encoder is illustrated in Figure 5. The Time Domain Bandwidth Extension (TDBWE) encoder extracts a fairly coarse parametric description from the pre-processed and downsampled higher-band signal $s_{HB}(n)$. This parametric description comprises *time envelope* and *frequency envelope* parameters. Refer to clauses 6.5.1 and 6.5.2 for the respective envelope computations and to 6.5.3 for a description of the parameter quantization scheme.

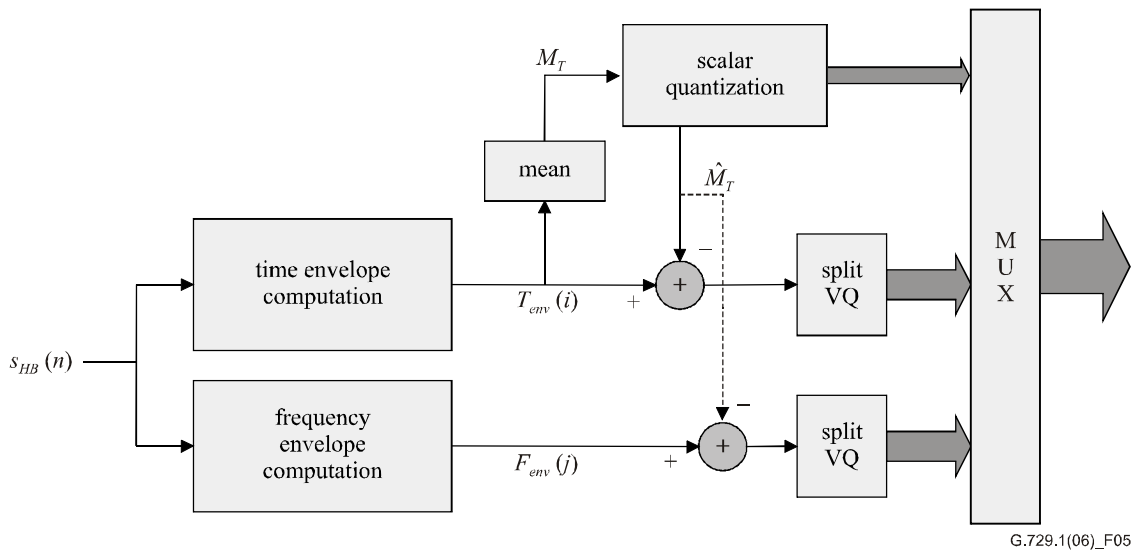


Figure 5 – High-level block diagram of the TDBWE encoder

6.5.1 Computation of the time envelope

The 20 ms input speech superframe $s_{HB}(n)$ is subdivided into 16 segments of length 1.25 ms each, i.e., each segment comprises 10 samples. The 16 time envelope parameters $T_{env}(i)$, $i = 0, \dots, 15$, are computed as logarithmic *subframe energies*:

$$T_{env}(i) = \frac{1}{2} \log_2 \left(\frac{1}{10} \sum_{n=0}^9 S_{HB}^2(n + i \cdot 10) \right), i = 0, \dots, 15 \quad (41)$$

6.5.2 Computation of the frequency envelope

For the computation of the 12 frequency envelope parameters $F_{env}(j)$, $j = 0, \dots, 11$, the signal $s_{HB}(n)$ is windowed by a slightly asymmetric analysis window $w_F(n)$. This window is 128 tap long (16 ms) and is constructed from the rising slope of a 144-tap Hanning window, followed by the falling slope of a 112-tap Hanning window:

$$w_F(n) = \begin{cases} \frac{1}{2} \left(1 - \cos \left(\frac{2\pi n}{144} \right) \right), & n = 0, \dots, 71 \\ \frac{1}{2} \left(1 - \cos \left(\frac{2\pi(n-16)}{113} \right) \right), & n = 72, \dots, 127 \end{cases}$$

The maximum of the window $w_F(n)$ is centered on the second 10 ms frame of the current superframe. The windowed signal over the second 10 ms frame is then given by:

$$s_{HB}^w(n) = s_{HB}(n) \cdot w_F(n + 31), \quad n = -31, \dots, 96 \quad (42)$$

The window $w_F(n)$ is constructed such that the frequency envelope computation has a lookahead of 16 samples (2 ms) and a lookback of 32 samples (4 ms). Figure 6 depicts the window function $w_F(n)$.

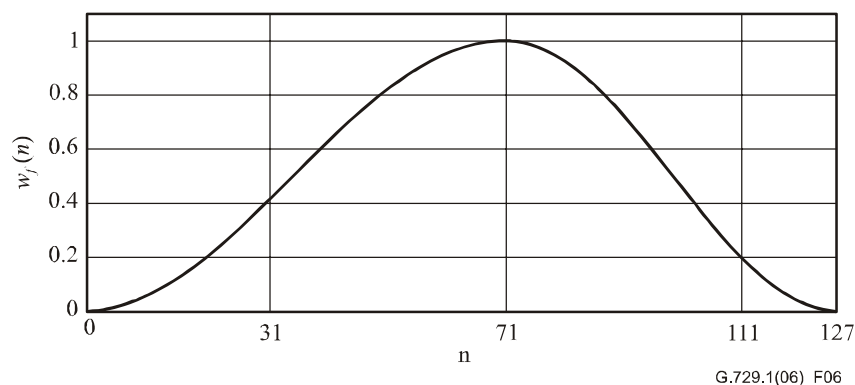


Figure 6 – Window for the frequency envelope computation

The windowed signal $s_{HB}^w(n)$ is transformed by FFT of length 64. With this FFT length, the *even* bins of the full length 128-tap FFT are computed using a polyphase structure:

$$S_{HB}^{ff}(k) = FFT_{64}(s_{HB}^w(n) + s_{HB}^w(n + 64)), \quad k = 0, \dots, 63, \quad n = -31, \dots, 32 \quad (43)$$

Finally, the frequency envelope parameter set is calculated as logarithmic weighted *sub-band energies* for 12 evenly spaced and equally wide overlapping sub-bands in the FFT domain:

$$F_{env}(j) = \frac{1}{2} \log_2 \left(\sum_{k=2j}^{2(j+1)} W_F(k - 2j) \cdot |S_{HB}^{ff}(k)|^2 \right), \quad j = 0, \dots, 11. \quad (44)$$

The j -th sub-band starts at the FFT bin of index $2j$ and spans a bandwidth of 3 FFT bins. This corresponds to the physical sub-band division as shown in Table 9. Note that, 0 Hz in this notation corresponds to a frequency of 4000 Hz in the original wideband frequency range.

Table 9 – Sub-band division for the frequency envelope

Sub-band index j	Sub-band frequency range [FFT bins] $k = 2j, \dots, 2(j+1)$	Sub-band frequency range [Hz] $f = \max\left(0, \frac{k-0.5}{64} \cdot 8000\text{Hz}\right) \dots \frac{k+0.5}{64} \cdot 8000\text{Hz}$
0	0,...,2	0 ... 312.5
1	2,...,4	187.5 ... 562.5
2	4,...,6	437.5 ... 812.5
3	6,...,8	687.5 ... 1062.5
4	8,...,10	937.5 ... 1312.5
5	10,...,12	1187.5 ... 1562.5
6	12,...,14	1437.5 ... 1812.5
7	14,...,16	1687.5 ... 2062.5
8	16,...,18	1937.5 ... 2312.5
9	18,...,20	2187.5 ... 2562.5
10	20,...,22	2437.5 ... 2812.5
11	22,...,24	2687.5 ... 3062.5

The physical *bandwidth* is Δf 375 Hz for each sub-band (but the first one). The frequency domain weighting window $W_F(k)$ is given as

$$W_F(0) = 0.5, W_F(1) = 1, \text{ and } W_F(2) = 0.5 \quad (45)$$

6.5.3 Quantization of TDBWE parameters

The TDBWE parameters $T_{env}(i)$, $i = 0, \dots, 15$, and $F_{env}(j)$, $j = 0, \dots, 11$, are quantized by mean-removed split vector quantization. First, a mean time envelope is calculated:

$$M_T = \frac{1}{16} \sum_{i=0}^{15} T_{env}(i) \quad (46)$$

The mean value M_T is then scalar quantized with 5 bits using uniform 3 dB steps in log domain. This quantization gives the quantized value $\hat{M}_T$. The quantized mean is then subtracted:

$$T_{env}^M(i) = T_{env}(i) - \hat{M}_T, \quad i = 0, \dots, 15 \quad (47)$$

and

$$F_{env}^M(j)_i = F_{env}(j) - \hat{M}_T, \quad j = 0, \dots, 11 \quad (48)$$

The mean-removed time envelope parameter set is split into two vectors of dimension 8

$$\mathbf{T}_{env,1} = (T_{env}^M(0), T_{env}^M(1), \dots, T_{env}^M(7)) \quad \text{and} \quad \mathbf{T}_{env,2} = (T_{env}^M(8), T_{env}^M(9), \dots, T_{env}^M(15)) \quad (49)$$

whereas the frequency envelope parameter is split into three vectors of dimension 4:

$$\begin{cases} \mathbf{F}_{env,1} = (F_{env}^M(0), F_{env}^M(1), F_{env}^M(2), F_{env}^M(3)) \\ \mathbf{F}_{env,2} = (F_{env}^M(4), F_{env}^M(5), F_{env}^M(6), F_{env}^M(7)) \\ \mathbf{F}_{env,3} = (F_{env}^M(8), F_{env}^M(9), F_{env}^M(10), F_{env}^M(11)) \end{cases} \quad (50)$$

Finally, vector quantization using pre-trained quantization tables is applied as shown in Table 10. Note that the vectors $\mathbf{T}_{env,1}$ and $\mathbf{T}_{env,2}$ share the same vector quantization codebooks to reduce storage requirements. The codebooks (or quantization tables) for $\mathbf{T}_{env,1}/\mathbf{T}_{env,2}$, $\mathbf{F}_{env,1}$, $\mathbf{F}_{env,2}$ and $\mathbf{F}_{env,3}$ have been generated by modifying generalized Lloyd-Max centroids such that a minimal distance between two centroids is verified. The codebook modification procedure consists in rounding Lloyd-Max centroids on a rectangular grid with a step size of 6 dB in log domain.

Table 10 – Time and frequency envelope quantization

Parameter vector	Quantized vector	Dimension	Number of allocated bits
M_T	$\hat{M}_T$	1	5
$\mathbf{T}_{env,1}$	$\hat{\mathbf{T}}_{env,1}$	8	7
$\mathbf{T}_{env,2}$	$\hat{\mathbf{T}}_{env,2}$	8	7
$\mathbf{F}_{env,1}$	$\hat{\mathbf{F}}_{env,1}$	4	5
$\mathbf{F}_{env,2}$	$\hat{\mathbf{F}}_{env,2}$	4	5
$\mathbf{F}_{env,3}$	$\hat{\mathbf{F}}_{env,3}$	4	4

6.5.4 Phase information for FEC

As mentioned in 6.3, the concealment/recovery parameters, which include signal classification information, energy, and phase information, are transmitted in different bitstream layers. The phase information, which consists of the position of the last glottal pulse in previous superframe, is transmitted in Layer 3 (14 kbit/s).

The phase control is particularly important while recovering after a lost segment of voiced speech since after a block of erased superframes, the decoder memories become desynchronized with the encoder memories. The transmitted phase information consists of the position and sign of the last glottal pulse in the previous superframe. This information is then used for the recovery after lost voiced onsets as described in 7.6.7. More importantly, this information is also used to resynchronize the excitation signal of erased superframes in order to improve the convergence in the correctly received consecutive superframes and reduce error propagation.

In the G.729EV decoder (in DEFAULT mode), one superframe delay is available at the decoder for the overlap-and-add operation in the MDCT reconstruction. Thus, when a single superframe is erased, the parameters of the future superframe are available (because of the extra superframe delay). In this case, the position and sign of the maximum pulse at the end of the erased superframe are available from the future superframe. Therefore the pitch excitation can be concealed in a way that the last maximum pulse is aligned with the position received in the future superframe as will be explained below.

The search of the maximum pulse is performed on a low-pass filtered LP residual, which is given by:

$$r^{lpf}(n) = 0.25r(n-1) + 0.5r(n) + 0.25r(n+1) \quad (51)$$

The position of the last glottal pulse τ is searched by looking for the sample with the maximum absolute amplitude among the T_0 last samples of the low-pass filtered residual $r^{lpf}(n)$ in the superframe, where T_0 is the rounded closed-loop pitch lag $T_2(2)$ of the last subframe: $T_0 = \text{int}(T_2(2))$. Note that τ is the position relative to the end of the superframe.

The position of the last glottal pulse τ is coded using 6 bits in the following manner. The precision used to encode the position of the first glottal pulse depends on T_0 . This is possible because this value is known both by the encoder and the decoder, and is not subject to error propagation after one or several superframe losses. If $T_0 < 64$, the position of the last glottal pulse relative to the end of the superframe is encoded directly with a precision of one sample. If $64 \leq T_0 < 128$, the position of the last glottal pulse relative to the end of the superframe is encoded with a precision of two

samples by using a simple integer division, i.e., $\tau/2$. If $T_0 \geq 128$, the position of the last glottal pulse relative to the end of the superframe is encoded with a precision of four samples by further dividing τ by 2.

The sign of the maximum absolute pulse amplitude is also transmitted with 1 bit. This gives a total of 7 bits for the phase information. The sign is important for phase resynchronization since the glottal pulse shape often contains two large pulses with opposite signs. Ignoring the sign may result in a small drift in the position and reduce the performance of the resynchronization procedure.

6.6 TDAC encoder (Layers 4 to 12)

The Time Domain Aliasing Cancellation (TDAC) encoder is illustrated in Figure 7. The TDAC encoder represents jointly two split MDCT spectra $D_{LB}^w(k)$ and $S_{HB}(k)$ by gain-shape vector quantization. The joint spectrum is divided into sub-bands. The gains in each sub-band define the spectral envelope. The shape of each sub-band is encoded by embedded spherical vector quantization using trained permutation codes.

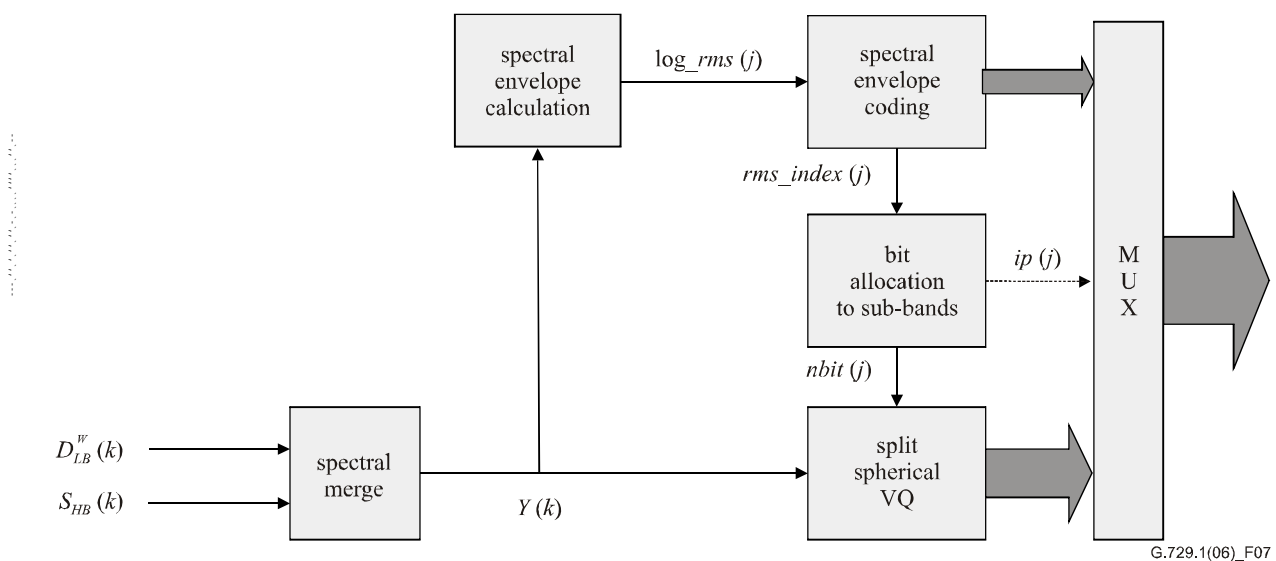


Figure 7 – High-level block diagram of the TDAC encoder

6.6.1 Energy information for FEC

The third concealment/recovery parameter is the energy information, which is computed based on the local synthesis signal at 12 kbit/s $\hat{s}_{enh}(n)$ at the encoder. The energy information is transmitted in Layer 4 (16 kbit/s). Thus if Layer 4 is received, this information can be used to improve the frame erasure concealment. Otherwise the energy is estimated at the decoder side.

The energy information E is the maximum sample energy for superframes classified as VOICED or ONSET, or the average energy per sample for other superframes. For VOICED or ONSET superframes, the maximum sample energy is computed pitch synchronously at the end of the superframe as follows:

$$E = \max_{n=160-t_E \dots 159} \hat{s}_{enh}^2(n) \quad (52)$$

where t_E is computed based on the closed-loop pitch lag $T_2(2)$ of the last subframe of the second frame. If $T_2(2) > 40$, $t_E = \text{round}(T_2(2))$ else $t_E = 2 \text{round}(T_2(2))$. For other classes, E is the average energy per sample of the second 10 ms frame, $t_E = 80$ and E is computed as:

$$E = \frac{1}{t_E} \sum_{n=160-t_E}^{159} \hat{s}_{enh}^2(n) \quad (53)$$

A 5-bit uniform quantizer is used in the range of 0 to 96 dB with a step size of 3.1 dB. The quantization index is given by:

$$i = \text{int} \left(\frac{10 \log_{10}(E + 0.001)}{3.1} \right) \quad (54)$$

where the index is bounded to $0 \leq i \leq 31$.

6.6.2 Perceptual weighting of the CELP difference signal

The difference $d_{LB}(n)$ between the embedded CELP encoder input $s(n)$ and the 12 kbit/s local synthesis $\hat{s}_{enh}(n)$ is processed by a perceptual weighting filter $W_{LB}(z)$ defined as:

$$W_{LB}(z) = \text{fac} \frac{\hat{A}(z/\gamma_1')}{\hat{A}(z/\gamma_2')} \quad (55)$$

where fac is a gain compensation given by:

$$\text{fac} = \frac{\left| \sum_{i=0}^{10} (-\gamma_2')^i \hat{a}_i \right|}{\left| \sum_{i=0}^{10} (-\gamma_1')^i \hat{a}_i \right|} \quad (56)$$

and $\hat{a}_i$ are the coefficients of the quantized linear-prediction filter $\hat{A}(z)_i$ obtained from the embedded CELP encoder with:

$$\hat{A}(z) = \hat{a}_0 + \hat{a}_1 z^{-1} + \dots + \hat{a}_{10} z^{-10} \quad (57)$$

The coefficients $\hat{a}_i$ are updated for every 5 ms subframe as explained in clauses 3.2.5 and 3.2.6/G.729. The parameters $\gamma_1' = 0.94$ and $\gamma_2' = 0.6$ are constant. Equivalently, the factor fac can be defined as the inverse gain of $\hat{A}(z/\gamma_1')/\hat{A}(z/\gamma_2')$ at Nyquist frequency of the lower band (4000 Hz):

$$\text{fac} = \frac{\left| \hat{A}(z/\gamma_2') \right|}{\left| \hat{A}(z/\gamma_1') \right|} \quad (58)$$

for $z = -1$. The gain compensation factor is required to guarantee the spectral continuity between the output $d_{LB}^w(n)$ of $W_{LB}(z)$ and the signal $s_{HB}(n)$ in the adjacent higher band.

The filter $W_{LB}(z)$ models the short-term inverse frequency masking curve and allows applying MDCT coding optimized for the mean-square error criterion. It also maps the difference signal $d_{LB}(n)$ into a weighted domain similar to the CELP target domain used at 8 and 12 kbit/s.

6.6.3 MDCT

The lower- and higher-band signals $d_{LB}^w(n)$ and $s_{HB}(n)$ are transformed into frequency domain by modified discrete cosine transform (MDCT) with a superframe length of 20 ms and a window length of 40 ms. The MDCT $D_{LB}^w(k)$ of the signal $d_{LB}^w(n)$ with a 40 ms sinusoidal windowing is given by:

$$D_{LB}^w(k) = \sqrt{\frac{2}{160}} \sum_{n=0}^{319} w_{TDAC}(n) \cos\left(\frac{\pi}{160}(n+80.5)(k+0.5)\right) d_{LB}^w(n), \quad k=0,\dots,159 \quad (59)$$

where $w_{TDAC}(n)$ is the analysis weighting window

$$w_{TDAC}(n) = \sin\left(\frac{\pi}{320}(n+0.5)\right), \quad n=0,\dots,319 \quad (60)$$

The computation of the 160 MDCT coefficients is done as follows.

First the real and imaginary parts of the complex numbers $z(n) = z_R(n) + j z_I(n)$, $n=0,\dots,79$, are calculated, with

$$\begin{aligned} z_R(n) &= w_{TDAC}(2n) d_{LB}^w(2n) - w_{TDAC}(159-2n) d_{LB}^w(159-2n) \\ z_I(n) &= w_{TDAC}(319-2n) d_{LB}^w(319-2n) + w_{TDAC}(160+2n) d_{LB}^w(160+2n) \end{aligned} \quad (61)$$

The complex signal $z(n)$ is then modified to $z'(n)$:

$$z'(n) = (W_{320})^n \times z(n), \quad n=0,\dots,79 \quad (62)$$

where $W_m = e^{2\pi j/m}$ denotes the m -th root of the unit.

The inverse FFT of $z'(n)$ is computed to obtain the coefficients $Z'(k)$, $k=0,\dots,79$. These coefficients are modified into:

$$Z(k) = (-1)^{k+1} \times (W_8)^{-1} \times (W_{1280})^{4k+1} \times Z'(k) \quad k=0,\dots,79 \quad (63)$$

where $Z_R(k)$ and $Z_I(k)$ are the real and imaginary parts of $Z(k)$. The MDCT coefficients are given by

$$\begin{cases} D_{LB}^w(2k) = Z_I(k) \\ D_{LB}^w(159-2k) = -Z_R(k) \end{cases} \quad \text{for } k=0,\dots,79 \quad (64)$$

The MDCT $S_{HB}(k)$ of $s_{HB}(n)$ is computed in the same way. The resulting MDCT spectra $D_{LB}^w(k)$ and $S_{HB}(k)$ of the lower and higher bands are merged into a full-band spectrum $Y(k)$:

$$[Y(0) Y(1) \dots Y(319)] = [D_{LB}^w(0) D_{LB}^w(1) \dots D_{LB}^w(159) S_{HB}(0) S_{HB}(1) \dots S_{HB}(159)] \quad (65)$$

Note that to maximize precision and avoid fixed-point arithmetic overflows, the signals $d_{LB}^w(n)$ and $s_{HB}(n)$ are normalized prior to MDCT. Since the TDAC encoder relies on two MDCTs, one in the lower band and one in the higher band, two normalization factors $norm_lo$ and $norm_hi$ are computed to scale the input. However, only one normalization factor is transmitted to the decoder with 4 bits:

$$norm_MDCT = \min(norm_lo, norm_hi) \quad (66)$$

6.6.4 Sub-band division

The MDCT coefficients in the 0-7000 Hz band are split into 18 sub-bands. Table 11 defines the sub-band boundaries and size. The j -th sub-band comprises $nb_coef(j)$ coefficients $Y(k)$ with $sb_bound(j) \leq k < sb_bound(j+1)$. The first 17 sub-bands comprise 16 coefficients (400 Hz), the last sub-band 8 coefficients (200 Hz).

Table 11 – Sub-band boundaries and number of coefficients per sub-band in the TDAC coder

j	$sb_bound(j)$	$nb_coef(j)$
0	0	16
1	16	16
2	32	16
3	48	16
4	64	16
5	80	16
6	96	16
7	112	16
8	128	16
9	144	16
10	160	16
11	176	16
12	192	16
13	208	16
14	224	16
15	240	16
16	256	16
17	272	8
18	280	—

6.6.5 Spectral envelope calculation

The spectral envelope is defined as the root mean square (rms) in log domain of the 18 sub-bands:

$$\log_rms(j) = \frac{1}{2} \log_2 \left[\frac{1}{nb_coef(j)} \sum_{k=sb_bound(j)}^{sb_bound(j+1)-1} Y(k)^2 + \epsilon_{rms} \right], \quad j = 0, \dots, 17 \quad (67)$$

where:

$$\epsilon_{rms} = 2^{-24}.$$

6.6.6 Spectral envelope coding

The spectral envelope is quantized with 5 bits by uniform scalar quantization and the resulting quantization indices are coded using a two-mode binary encoder. The 5-bit quantization consists in computing the indices $rms_index(j)$, $j=0, \dots, 17$, as follows:

$$rms_index(j) = round(2 \log_rms(j)) \quad (68)$$

with the restriction

$$-11 \leq rms_index(j) \leq +20 \quad (69)$$

i.e., the indices are limited by -11 and $+20$ (32 possible values). The resulting quantized full-band envelope is then divided into two subvectors:

- lower-band spectral envelope: $(rms_index(0), rms_index(1), \dots, rms_index(9))$ and
- higher-band spectral envelope: $(rms_index(10), rms_index(11), \dots, rms_index(17))$.

These two subvectors are coded separately using a two-mode lossless encoder which switches adaptively between differential Huffman coding (mode 0) and direct natural binary coding (mode 1). Differential Huffman coding is used to minimize the average number of bits, whereas direct natural binary coding is used to limit the worst-case number of bits as well as to correctly encode the envelope of signals which are saturated by differential Huffman coding (e.g., sinusoids). One bit is used to indicate the selected mode to the spectral envelope decoder.

Encoding of the lower-band envelope

The differential Huffman encoder (mode 0) comprises four steps:

Step 0: Natural binary coding of first (absolute) index $rms_index(0)$. The index $rms_index(0)$, which is comprised in $[-11, +20]$, is coded with 5 bits.

Step 1: Computation of differential indices and saturation flag *satur*

$$diff_index(j) = rms_index(j) - rms_index(j-1), \quad j = 1, \dots, 9 \quad (70)$$

$$satur = \begin{cases} 0 & \text{if } |diff_index(j)| \leq 12 \text{ for } j = 1, \dots, 9 \\ 1 & \text{otherwise} \end{cases} \quad (71)$$

The binary flag *satur* is used to detect cases when $diff_index(j)$ is outside $[-12, +12]$.

Step 2: If *satur* = 0, differential Huffman coding of differential indices and *mode* is set to 0.

The differential indices $diff_index(j)$ for $j=1 \dots 9$ are Huffman coded by table-lookup. Table 12 describes the related Huffman codes and code length.

Step 3: If *satur* = 1 or the number of bits used by differential Huffman coding is greater than 45, natural binary encoding of $rms_index(1) \dots rms_index(9)$ as in Step 0 is applied and *mode* is set to 1.