



TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(07/2003)

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

Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)

ITU-T Recommendation G.722.2

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
TESTING EQUIPMENTS	G.500–G.599
TRANSMISSION MEDIA 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
DIGITAL TERMINAL EQUIPMENTS	G.7000–G.7999
DIGITAL NETWORKS	G.8000–G.8999

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

ITU-T Recommendation G.722.2

Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)

Summary

This Recommendation describes the high quality Adaptive Multi-Rate Wideband (AMR-WB) encoder and decoder that is primarily intended for 7 kHz bandwidth speech signals. AMR-WB operates at a multitude of bit rates ranging from 6.6 kbit/s to 23.85 kbit/s. The bit rate may be changed at any 20-ms frame boundary.

Annex C includes an integrated C source code software package which contains the implementation of the G.722.2 encoder and decoder and its Annexes A and B and Appendix I. A set of digital test vectors for developers is provided in Annex D. These test vectors are a verification tool providing an indication of success in implementing this codec.

G.722.2 AMR-WB is the same codec as the 3GPP AMR-WB. The corresponding 3GPP specifications are TS 26.190 for the speech codec and TS 26.194 for the Voice Activity Detector.

Source

ITU-T Recommendation G.722.2 was approved on 29 July 2003 by ITU-T Study Group 16 (2001-2004) under the ITU-T Recommendation A.8 procedure.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. 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, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

© ITU 2004

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	Normative references.....	2
3	Definitions, symbols and abbreviations.....	2
3.1	Definitions	2
3.2	Symbols	3
3.3	Abbreviations	7
4	Outline description.....	8
4.1	Functional description of audio parts	8
4.2	Preparation of speech samples.....	8
4.3	Principles of the adaptive multi-rate wideband speech encoder	8
4.4	Principles of the adaptive multi-rate speech decoder	13
4.5	Sequence and subjective importance of encoded parameters.....	14
5	Functional description of the encoder.....	14
5.1	Preprocessing.....	14
5.2	Linear prediction analysis and quantization	15
5.3	Perceptual weighting	20
5.4	Open-loop pitch analysis	20
5.5	Impulse response computation	22
5.6	Target signal computation	22
5.7	Adaptive codebook.....	23
5.8	Algebraic codebook.....	25
5.9	Quantization of the adaptive and fixed codebook gains.....	36
5.10	Memory update.....	37
5.11	High-band gain generation	37
6	Functional description of the decoder.....	37
6.1	Decoding and speech synthesis	38
6.2	High-pass filtering, up-scaling and interpolation	40
6.3	High frequency band	41
7	Detailed bit allocation of the adaptive multi-rate wideband codec	43
8	Homing sequences.....	51
8.1	Functional description	51
8.2	Definitions	52
8.3	Encoder homing.....	52
8.4	Decoder homing	52

	Page
9 Voice Activity Detector (VAD).....	53
9.1 VAD symbols	53
9.2 Functional description	55
10 Mandatory AMR-WB codec modes for the speech telephony service in 3GPP	62
BIBLIOGRAPHY	64

iv

ITU-T Recommendation G.722.2

Wideband coding of speech at around 16 kbit/s using Adaptive Multi-Rate Wideband (AMR-WB)

1 Scope

This Recommendation describes the detailed mapping from input blocks of 320 speech samples in 16-bit uniform PCM format to encoded blocks of 132, 177, 253, 285, 317, 365, 397, 461 and 477 bits and from encoded blocks of 132, 177, 253, 285, 317, 365, 397, 461 and 477 bits to output blocks of 320 reconstructed speech samples. The sampling rate is 16 000 samples/s leading to a bit rate for the encoded bit stream of 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 or 23.85 kbit/s. The coding scheme for the multi-rate coding modes is the so-called Algebraic Code Excited Linear Prediction Coder, hereafter referred to as ACELP. The multi-rate wideband ACELP coder is referred to as AMR-WB. The codec described in this Recommendation also utilizes an integrated Voice Activity Detector (VAD).

The foreseen applications for this Recommendation are the following: Voice over IP (VoIP) and Internet applications, Mobile Communications, PSTN applications, ISDN wideband telephony, ISDN videotelephony and video-conferencing.

In addition to the algorithm specified in the main body of this Recommendation, Annexes A and B and Appendix I provide supplemental functionalities allowing interoperability with GSM and 3GPP wireless systems. These functionalities have originally been developed for these systems, but their use is not limited to mobile applications. Annexes D and E describe test vectors and frame structure respectively. These annexes may be implemented independently of this main body specification according to the different requirements of systems deploying the AMR-WB algorithm:

- Annex A describes comfort noise aspects for use of the AMR-WB algorithm in source-controlled rate operation. The implementation of Annex A is essential for interoperability with GSM and 3GPP wireless systems.
- Annex B describes the source-controlled rate operation for the AMR-WB algorithm. The implementation of Annex B is essential for interoperability with GSM and 3GPP wireless systems.
- Annex D describes the digital test sequences, which are a verification tool providing an indication of success in implementing the AMR-WB codec.
- Annex E describes the recommended frame structure for use with the different modes of operation for the AMR-WB algorithm.
- Appendix I describes an example solution for error concealment of erroneous or lost AMR-WB frames.

For better usability, the ANSI-C code with the low-level description of all these functionalities have been grouped into a single annex, Annex C. Should there be any discrepancy between the descriptions in any of the different parts of this Recommendation and the implementation of such descriptions in Annex C, the descriptions in Annex C shall prevail.

In clause 8 a specific reset procedure, called codec homing, is described. This is a useful feature for bringing the codec into a known initial state (e.g., for testing purposes). Clause 9 specifies the Voice Activity Detector (VAD) used in this codec as well as in the source controlled rate operation (DTX) specified in Annex B. Clause 10 provides information on minimum requirements for support of AMR-WB modes in the 3GPP speech telephony service.

2 Normative 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.

- [1] ITU-T Recommendation G.722 (1988), *7 kHz audio-coding within 64 kbit/s*.
- [2] RFC 3267 (2002), *Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs*.

3 Definitions, symbols and abbreviations

3.1 Definitions

This Recommendation defines the following terms:

3.1.1 adaptive codebook: The adaptive codebook contains excitation vectors that are adapted for every subframe. The adaptive codebook is derived from the long-term filter state. The lag value can be viewed as an index into the adaptive codebook.

3.1.2 algebraic codebook: A fixed codebook where algebraic code is used to populate the excitation vectors (innovation vectors). The excitation contains a small number of non-zero pulses with predefined interlaced sets of potential positions. The amplitudes and positions of the pulses of the k th excitation codevector can be derived from its index k through a rule requiring no or minimal physical storage, in contrast with stochastic codebooks, whereby the path from the index to the associated codevector involves look-up tables.

3.1.3 anti-sparseness processing: An adaptive post-processing procedure applied to the fixed codebook vector in order to reduce perceptual artifacts from a sparse fixed codebook vector.

3.1.4 closed-loop pitch analysis: This is the adaptive codebook search, i.e. a process of estimating the pitch (lag) value from the weighted input speech and the long-term filter state. In the closed-loop search, the lag is searched using error minimization loop (analysis-by-synthesis). In the adaptive multi-rate wideband codec, closed-loop pitch search is performed for every subframe.

3.1.5 direct form coefficients: One of the formats for storing the short-term filter parameters. In the adaptive multi-rate wideband codec, all filters which are used to modify speech samples use direct form coefficients.

3.1.6 fixed codebook: The fixed codebook contains excitation vectors for speech synthesis filters. The contents of the codebook are non-adaptive (i.e., fixed). In the adaptive multi-rate wideband codec, the fixed codebook is implemented using an algebraic codebook.

3.1.7 fractional lags: A set of lag values having subsample resolution. In the adaptive multirate wideband codec a sub-sample resolution of 1/4th or 1/2nd of a sample is used.

3.1.8 frame: A time interval equal to 20 ms (320 samples at a 16-kHz sampling rate).

3.1.9 immittance spectral frequencies: (see Immittance Spectral Pair)

3.1.10 immittance spectral pair: Transformation of LPC parameters. Immittance Spectral Pairs are obtained by decomposing the inverse filter transfer function $A(z)$ to a set of two transfer functions, one having even symmetry and the other having odd symmetry. The Immittance Spectral

Pairs (also called Immittance Spectral Frequencies) are the roots of these polynomials on the z-unit circle.

3.1.11 integer lags: A set of lag values having whole sample resolution.

3.1.12 interpolating filter: An FIR filter used to produce an estimate of subsample resolution samples, given an input sampled with integer sample resolution. In this implementation, the interpolating filter has low-pass filter characteristics. Thus the adaptive codebook consists of the low-pass filtered interpolated past excitation.

3.1.13 inverse filter: This filter removes the short-term correlation from the speech signal. The filter models an inverse frequency response of the vocal tract.

3.1.14 lag: The long-term filter delay. This is typically the true pitch period, or its multiple or submultiple.

3.1.15 LP analysis window: For each frame, the short-term filter coefficients are computed using the high-pass filtered speech samples within the analysis window. In the adaptive multi-rate wideband codec, the length of the analysis window is always 384 samples. For all the modes, a single asymmetric window is used to generate a single set of LP coefficients. The 5-ms look-ahead is used in the analysis

3.1.16 LP coefficients: Linear Prediction (LP) coefficients (also referred to as Linear Predictive Coding (LPC) coefficients) is a generic descriptive term for the short-term filter coefficients.

3.1.17 mode: When used alone, refers to the source codec mode, i.e., to one of the source codecs employed in the AMR-WB codec

3.1.18 open-loop pitch search: A process of estimating the near optimal lag directly from the weighted speech input. This is done to simplify the pitch analysis and confine the closed-loop pitch search to a small number of lags around the open-loop estimated lags. In the adaptive multi-rate wideband codec, an open-loop pitch search is performed in every other subframe.

3.1.19 residual: The output signal resulting from an inverse filtering operation.

3.1.20 short-term synthesis filter: This filter introduces short-term correlation into the excitation signal which models the impulse response of the vocal tract.

3.1.21 perceptual weighting filter: This filter is employed in the analysis-by-synthesis search of the codebooks. The filter exploits the noise masking properties of the formants (vocal tract resonances) by weighting the error less in regions near the formant frequencies and more in regions away from them.

3.1.22 subframe: A time interval equal to 5 ms (80 samples at 16-kHz sampling rate).

3.1.23 vector quantization: A method of grouping several parameters into a vector and quantizing them simultaneously.

3.1.24 zero input response: The output of a filter due to past inputs, i.e., due to the present state of the filter, given that an input of zeros is applied.

3.1.25 zero state response: The output of a filter due to the present input, given that no past inputs have been applied, i.e., given that the state information in the filter is all zeros.

3.2 Symbols

This Recommendation uses the following symbols:

$A(z)$ The inverse filter with unquantized coefficients

$\hat{A}(z)$ The inverse filter with quantized coefficients

$H(z) = \frac{1}{\hat{A}(z)}$	The speech synthesis filter with quantized coefficients
a_i	The unquantized linear prediction parameters (direct form coefficients)
$\hat{a}_i$	The quantified linear prediction parameters
m	The order of the LP model
$W(z)$	The perceptual weighting filter (unquantized coefficients)
γ_1	The perceptual weighting factor
T	The integer pitch lag nearest to the closed-loop fractional pitch lag of the subframe
β	The adaptive prefilter coefficient (the quantified pitch gain)
$H_{h1}(z)$	Preprocessing high-pass filter
$w(n)$	LP analysis window
L_1	Length of the first part of the LP analysis window $w(n)$
L_2	Length of the second part of the LP analysis window $w(n)$
$r(k)$	The autocorrelations of the windowed speech $s'(n)$
$w_{lag}(i)$	Lag window for the autocorrelations (60-Hz bandwidth expansion)
f_0	The bandwidth expansion in Hz
f_s	The sampling frequency in Hz
$r'(k)$	The modified (bandwidth expanded) autocorrelations
$E(i)$	The prediction error in the i th iteration of the Levinson algorithm
k_i	The i th reflection coefficient
$a_j^{(i)}$	The j th direct form coefficient in the i th iteration of the Levinson algorithm
$F_1'(z)$	Symmetric ISF polynomial
$F_2'(z)$	Antisymmetric ISF polynomial
$F_1(z)$	Polynomial $F_1'(z)$
$F_2(z)$	Polynomial $F_2'(z)$ with roots $z = 1$ and $z = -1$ eliminated
q_i	The immittance spectral pairs (ISPs) in the cosine domain
$\mathbf{q}$	An ISP vector in the cosine domain
$\hat{\mathbf{q}}_i^{(n)}$	The quantified ISP vector at the i th subframe of the frame n
ω_i	The immittance spectral frequencies (ISFs)
$T_m(x)$	A m th order Chebyshev polynomial
$f_1(i), f_2(i)$	The coefficients of the polynomials $F_1(z)$ and $F_2(z)$
$f_1'(i), f_2'(i)$	The coefficients of the polynomials $F_1'(z)$ and $F_2'(z)$
$f(i)$	The coefficients of either $F_1(z)$ or $F_2(z)$

$C(x)$	Sum polynomial of the Chebyshev polynomials
x	Cosine of angular frequency ω
λ_k	Recursion coefficients for the Chebyshev polynomial evaluation
f_i	The immittance spectral frequencies (ISFs) in Hz
$\mathbf{f}^t = [f_1 \ f_2 \ \cdots \ f_{16}]$	The vector representation of the ISFs in Hz
$\mathbf{z}(n)$	The mean-removed ISF vector at frame n
$\mathbf{r}(n)$	The ISF prediction residual vector at frame n
$\mathbf{p}(n)$	The predicted ISF vector at frame n
$\hat{\mathbf{r}}(n-1)$	The quantified residual vector at the past frame
$\hat{\mathbf{r}}_i^k$	The quantified ISF subvector i at quantization index k
d_i	The distance between the immittance spectral frequencies f_{i+1} and f_{i-1}
$h(n)$	The impulse response of the weighted synthesis filter
$H(z)W(z)$	The weighted synthesis filter
T_1	The integer nearest to the fractional pitch lag of the previous (1st or 3rd) subframe
$s'(n)$	The windowed speech signal
$s_w(n)$	The weighted speech signal
$\hat{s}(n)$	Reconstructed speech signal
$x(n)$	The target signal for adaptive codebook search
$x_2(n), \mathbf{x}_2^t$	The target signal for algebraic codebook search
$res_{LP}(n)$	The LP residual signal
$c(n)$	The fixed codebook vector
$v(n)$	The adaptive codebook vector
$y(n) = v(n) * h(n)$	The filtered adaptive codebook vector
$y_k(n)$	The past filtered excitation
$u(n)$	The excitation signal
$\hat{u}'(n)$	The gain-scaled emphasized excitation signal
T_{op}	The best open-loop lag
t_{min}	Minimum lag search value
t_{max}	Maximum lag search value
$R(k)$	Correlation term to be maximized in the adaptive codebook search
$R(k)_t$	The interpolated value of $R(k)$ for the integer delay k and fraction t
A_k	Correlation term to be maximized in the algebraic codebook search at index k
C_k	The correlation in the numerator of A_k at index k
E_{Dk}	The energy in the denominator of A_k at index k

$\mathbf{d} = \mathbf{H}^t \mathbf{x}_2$	The correlation between the target signal $x_2(n)$ and the impulse response $h(n)$, i.e., backward filtered target
$\mathbf{H}$	The lower triangular Toeplitz convolution matrix with diagonal $h(0)$ and lower diagonals $h(1), \dots, h(63)$
$\Phi = \mathbf{H}^t \mathbf{H}$	The matrix of correlations of $h(n)$
$d(n)$	The elements of the vector $\mathbf{d}$
$\phi(i, j)$	The elements of the symmetric matrix Φ
c_k	The innovation vector
C	The correlation in the numerator of A_k
m_i	The position of the i th pulse
ϑ_i	The amplitude of the i th pulse
N_p	The number of pulses in the fixed codebook excitation
E_D	The energy in the denominator of A_k
$res_{LTP}(n)$	The normalized long-term prediction residual
$b(n)$	The signal used for presetting the signs in algebraic codebook search
$s_b(n)$	The sign signal for the algebraic codebook search
$d'(n)$	Sign extended backward filtered target
$\phi'(i, j)$	The modified elements of the matrix Φ , including sign information
$\mathbf{z}', z(n)$	The fixed codebook vector convolved with $h(n)$
$E(n)$	The mean-removed innovation energy (in dB)
$\bar{E}$	The mean of the innovation energy
$\tilde{E}(n)$	The predicted energy
$[b_1 \ b_2 \ b_3 \ b_4]$	The MA prediction coefficients
$\hat{R}(k)$	The quantified prediction error at subframe k
E_I	The mean innovation energy
$R(n)$	The prediction error of the fixed-codebook gain quantization
E_Q	The quantization error of the fixed-codebook gain quantization
$e(n)$	The states of the synthesis filter $1/\hat{A}(z)$
$e_w(n)$	The perceptually weighted error of the analysis-by-synthesis search
η	The gain scaling factor for the emphasized excitation
g_c	The fixed-codebook gain
g'_c	The predicted fixed-codebook gain
$\hat{g}_c$	The quantified fixed codebook gain
g_p	The adaptive codebook gain
$\hat{g}_p$	The quantified adaptive codebook gain

$\gamma_{gc} = g_c / g'_c$	A correction factor between the gain g_c and the estimated one g'_c
$\hat{\gamma}_{gc}$	The optimum value for γ_{gc}
γ_{sc}	Gain scaling factor

3.3 Abbreviations

This Recommendation uses the following abbreviations:

ACELP	Algebraic Code-Excited Linear Prediction
AGC	Adaptive Gain Control
AMR	Adaptive Multi-rate
AMR-WB	Adaptive Multi-rate Wideband
CELP	Code Excited Linear Prediction
FIR	Finite Impulse Response
GERAN	GSM EDGE Radio Access Network
GMSK	Gaussian Modified Shift Keying
ISF	Immittance Spectral Frequency
ISP	Immittance Spectral Pair
ISPP	Interleaved Single-Pulse Permutation
LP	Linear Prediction
LPC	Linear Predictive Coding
LTP	Long-Term Predictor (or Long-Term Prediction)
MA	Moving Average
O-TCH/F	Octal TCH/Full rate: a 68.4-kbit/s gross bit-rate radio channel of GERAN-8PSK
O-TCH/H	Octal-TCH/Half rate: a 32.4-kbit/s gross bit-rate radio channel of GERAN-8PSK
S-MSVQ	Split-MultiStage Vector Quantization
TCH	Traffic CHannel (dedicated radio channel for speech or data)
TCH/F	TCH/Full rate: a 22.8-kbit/s gross bit-rate radio channel of GERAN-GMSK
TCH/H	TCH/Half rate: a 11.4-kbit/s gross bit-rate radio channel of GERAN-GMSK
TFO	Tandem Free Operation (Tandem Free is achieved using in-band signalling after call set-up)
TrFO	Transcoder Free Operation (Tandem Free is achieved using out-of-band signalling before call set-up; the transcoders are in principle not in the communication path)
UTRAN	UMTS Terrestrial Radio Access Network
WB	Wideband
8PSK	8 Phase Shift Keying

4 Outline description

This Recommendation is structured as follows:

- Clause 4.1 contains a functional description of the audio parts including the A/D and D/A functions.
- Clause 4.2 describes input format for the AMR-WB encoder and the output format for the AMR-WB decoder.
- Clauses 4.3 and 4.4 present a simplified description of the principles of the AMR-WB codec encoding and decoding process respectively.
- In clause 4.5, the sequence and subjective importance of encoded parameters are given.
- Clause 5 presents the functional description of the AMR-WB codec encoding, whereas clause 6 describes the decoding procedures.
- In clause 7, the detailed bit allocation of the AMR-WB codec is tabulated.
- Clause 8 describes the homing operation.

4.1 Functional description of audio parts

The analogue-to-digital and digital-to-analogue conversion will, in principle, comprise the following elements:

- 1) Analogue to uniform digital PCM [1]
 - microphone;
 - input level adjustment device;
 - input anti-aliasing filter;
 - sample-hold device sampling at 16 kHz;
 - analogue-to-uniform digital conversion to 14-bit representation.

The uniform format shall be represented in two's complement.

- 2) Uniform digital PCM to analogue [1]
 - conversion from 14-bit/16-kHz uniform PCM to analogue;
 - a hold device;
 - reconstruction filter including $x/\sin(x)$ correction;
 - output level adjustment device;
 - earphone or loudspeaker.

In the terminal equipment, the A/D function may be achieved

- by direct conversion to 14-bit uniform PCM format;

For the D/A operation, the inverse operations take place.

4.2 Preparation of speech samples

The encoder is fed with data comprising samples with a resolution of 14 bits left justified in a 16-bit word. The decoder outputs data in the same format. Outside the speech codec, further processing must be applied if the traffic data occurs in a different representation.

4.3 Principles of the adaptive multi-rate wideband speech encoder

The AMR-WB codec consists of nine source codecs with bit rates of 23.85, 23.05, 19.85, 18.25, 15.85, 14.25, 12.65, 8.85 and 6.60 kbit/s.

The codec is based on the code excited linear prediction (CELP) coding model. The input signal is pre-emphasized using the filter $H_{\text{pre-emph}}(z) = 1 - \mu z^{-1}$. The CELP model is then applied to the pre-

emphasized signal. A 16th-order linear prediction (LP), or short-term, synthesis filter is used which is given by:

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

where $\hat{a}_i, i = 1, \dots, m$ are the (quantized) linear prediction (LP) parameters, and $m = 16$ is the predictor order. The long-term, or pitch, synthesis filter is usually given by:

$$\frac{1}{B(z)} = \frac{1}{1 - g_p z^{-T}} \quad (2)$$

where T is the pitch delay and g_p is the pitch gain. The pitch synthesis filter is implemented using the so-called adaptive codebook approach.

The CELP speech synthesis model is shown in Figure 1. In this model, the excitation signal at the input of the short-term LP synthesis filter is constructed by adding two excitation vectors from adaptive and fixed (innovative) codebooks. The speech is synthesized by feeding the two properly chosen vectors from these codebooks through the short-term synthesis filter. The optimum excitation sequence in a codebook is chosen using an analysis-by-synthesis search procedure in which the error between the original and synthesized speech is minimized according to a perceptually weighted distortion measure.

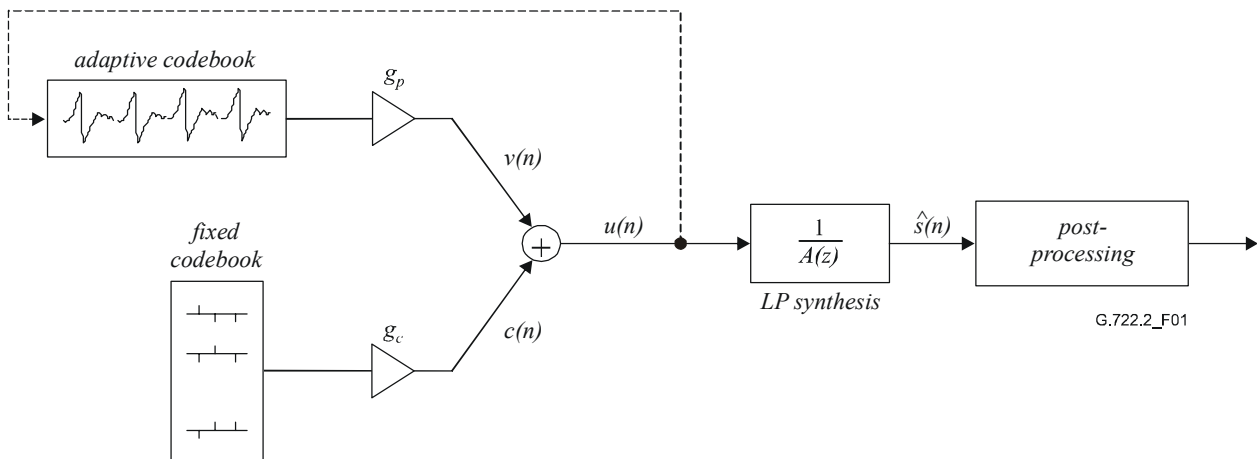


Figure 1/G.722.2 – Simplified block diagram of the CELP synthesis model

The perceptual weighting filter used in the analysis-by-synthesis search technique is given by:

$$W(z) = A(z/\gamma_1) H_{\text{de-emph}}(z) \quad (3)$$

where $A(z)$ is the unquantized LP filter, $H_{\text{de-emph}} = \frac{1}{1 - 0.68z^{-1}}$, and $\gamma_1 = 0.92$ is the perceptual weighting factor. The weighting filter uses the unquantized LP parameters.

The encoder performs the analysis of the LPC, LTP and fixed codebook parameters at 12.8-kHz sampling rate. The coder operates on speech frames of 20 ms. At each frame, the speech signal is analysed to extract the parameters of the CELP model (LP filter coefficients, adaptive and fixed codebooks' indices and gains). In addition to these parameters, high-band gain indices are computed in 23.85-kbit/s mode. These parameters are encoded and transmitted. At the decoder, these parameters are decoded and speech is synthesized by filtering the reconstructed excitation signal through the LP synthesis filter.

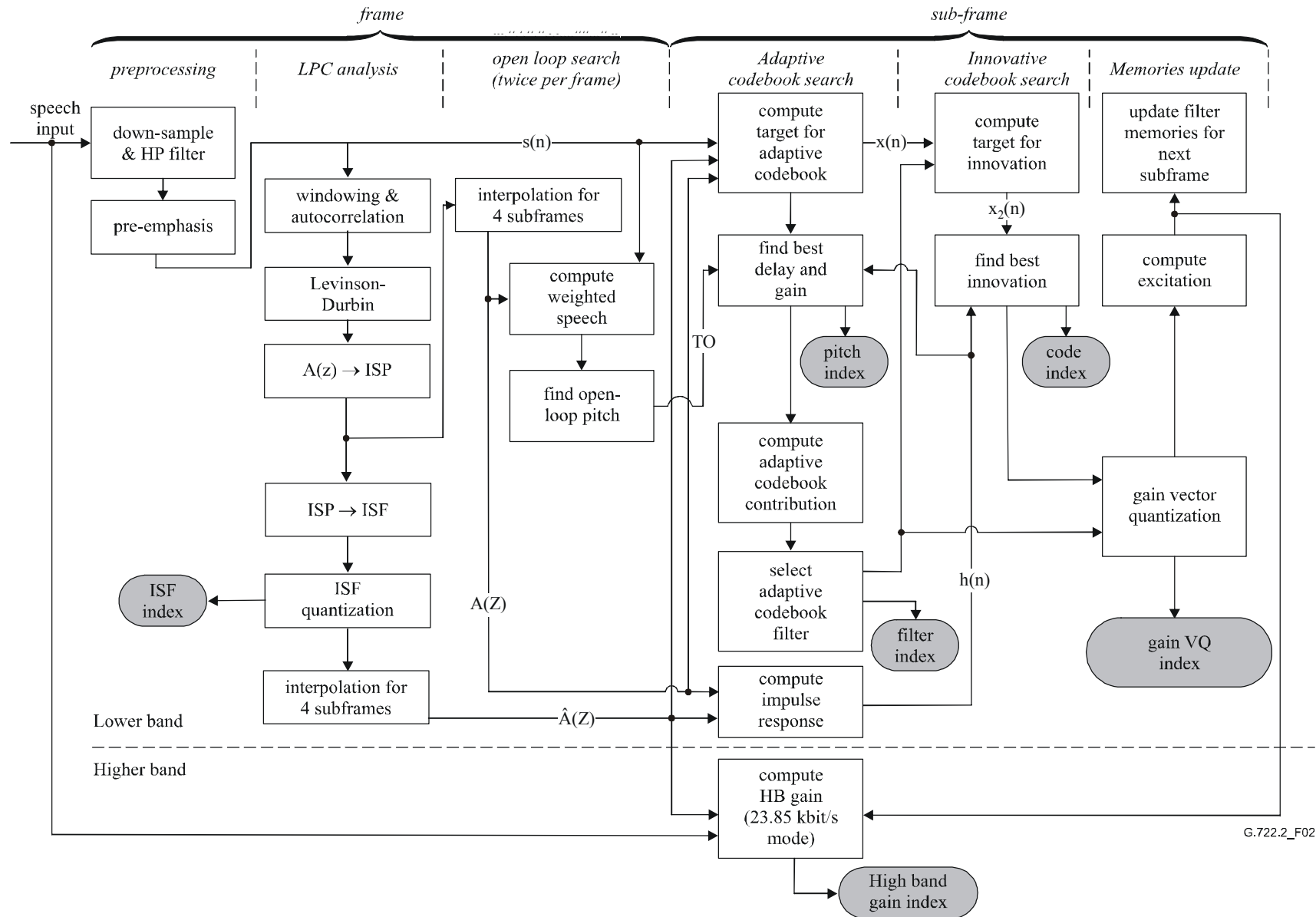


Figure 2/G.722.2 – Detailed block diagram of the ACELP encoder

The signal flow at the encoder is shown in Figure 2. After decimation, high-pass and pre-emphasis filtering is performed. LP analysis is performed once per frame. The set of LP parameters is converted to immittance spectrum pairs (ISP) and vector quantized using split-multistage vector quantization (S-MSVQ). The speech frame is divided into 4 subframes of 5 ms each (64 samples at 12.8-kHz sampling rate). The adaptive and fixed codebook parameters are transmitted every subframe. The quantized and unquantized LP parameters or their interpolated versions are used depending on the subframe. An open-loop pitch lag is estimated in every other subframe or once per frame based on the perceptually weighted speech signal.

The following operations are then repeated for each subframe:

- The target signal $x(n)$ is computed by filtering the LP residual through the weighted synthesis filter $W(z)H(z)$ with the initial states of the filters having been updated by filtering the error between LP residual and excitation (this is equivalent to the common approach of subtracting the zero input response of the weighted synthesis filter from the weighted speech signal).
- The impulse response $h(n)$ of the weighted synthesis filter is computed.
- Closed-loop pitch analysis is then performed (to find the pitch lag and gain), using the target $x(n)$ and impulse response $h(n)$, by searching around the open-loop pitch lag. Fractional pitch with 1/4th or 1/2nd of a sample resolution (depending on the mode and the pitch lag value) is used. The interpolating filter in fractional pitch search has low-pass frequency response. Further, there are two potential low-pass characteristics in the adaptive codebook and this information is encoded with 1 bit.
- The target signal $x(n)$ is updated by removing the adaptive codebook contribution (filtered adaptive codevector), and this new target, $x_2(n)$, is used in the fixed algebraic codebook search (to find the optimum innovation).
- The gains of the adaptive and fixed codebook are vector quantified with 6 or 7 bits (with moving average (MA) prediction applied to the fixed codebook gain).
- Finally, the filter memories are updated (using the determined excitation signal) for finding the target signal in the next subframe.

The bit allocation of the AMR-WB codec modes is shown in Table 1. In each 20-ms speech frame, 132, 177, 253, 285, 317, 365, 397, 461 and 477 bits are produced, corresponding to a bit rate of 6.60, 8.85, 12.65, 14.25, 15.85, 18.25, 19.85, 23.05 or 23.85 kbit/s. More detailed bit allocation among the codec parameters is given in Tables 12a-12i. Note that the most significant bits (MSB) are always sent first.

Table 1/G.722.2 – Bit allocation of the AMR-WB coding algorithm for 20-ms frame

Mode	Parameter	1st subframe	2nd subframe	3rd subframe	4th subframe	Total per frame
23.85 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	88	88	88	88	352
	Codebook gain	7	7	7	7	28
	HB-energy	4	4	4	4	16
	Total	477				

Table 1/G.722.2 – Bit allocation of the AMR-WB coding algorithm for 20-ms frame

Mode	Parameter	1st subframe	2nd subframe	3rd subframe	4th subframe	Total per frame
23.05 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	88	88	88	88	352
	Gains	7	7	7	7	28
	Total					461
19.85 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	72	72	72	72	288
	Codebook gain	7	7	7	7	28
	Total					397
18.25 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	64	64	64	64	256
	Gains	7	7	7	7	28
	Total					365
15.85 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	52	52	52	52	208
	Gains	7	7	7	7	28
	Total					317
14.25 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	44	44	44	44	176
	Gains	7	7	7	7	28
	Total					285

Table 1/G.722.2 – Bit allocation of the AMR-WB coding algorithm for 20-ms frame

Mode	Parameter	1st subframe	2nd subframe	3rd subframe	4th subframe	Total per frame
12.65 kbit/s	VAD-flag					1
	ISP					46
	LTP-filtering	1	1	1	1	4
	Pitch delay	9	6	9	6	30
	Algebraic code	36	36	36	36	144
	Gains	7	7	7	7	28
	Total					253
8.85 kbit/s	VAD-flag					1
	ISP					46
	Pitch delay	8	5	8	5	26
	Algebraic code	20	20	20	20	80
	Gains	6	6	6	6	24
	Total					177
6.60 kbit/s	VAD-flag					1
	ISP					36
	Pitch delay	8	5	5	5	23
	Algebraic code	12	12	12	12	48
	Gains	6	6	6	6	24
	Total					132

4.4 Principles of the adaptive multi-rate speech decoder

The signal flow at the decoder is shown in Figure 3. At the decoder, the transmitted indices are extracted from the received bitstream. The indices are decoded to obtain the coder parameters at each transmission frame. These parameters are the ISP vector, the 4 fractional pitch lags, the 4 LTP filtering parameters, the 4 innovative codevectors, and the 4 sets of vector quantized pitch and innovative gains. In 23.85-kbit/s mode, also high-band gain index is decoded. The ISP vector is converted to the LP filter coefficients and interpolated to obtain LP filters at each subframe. Then, at each 64-sample subframe:

- the excitation is constructed by adding the adaptive and innovative codevectors scaled by their respective gains;
- the 12.8-kHz speech is reconstructed by filtering the excitation through the LP synthesis filter;
- the reconstructed speech is de-emphasized.

Finally, the reconstructed speech is upsampled to 16 kHz and high-band speech signal is added to the frequency band from 6 kHz to 7 kHz.

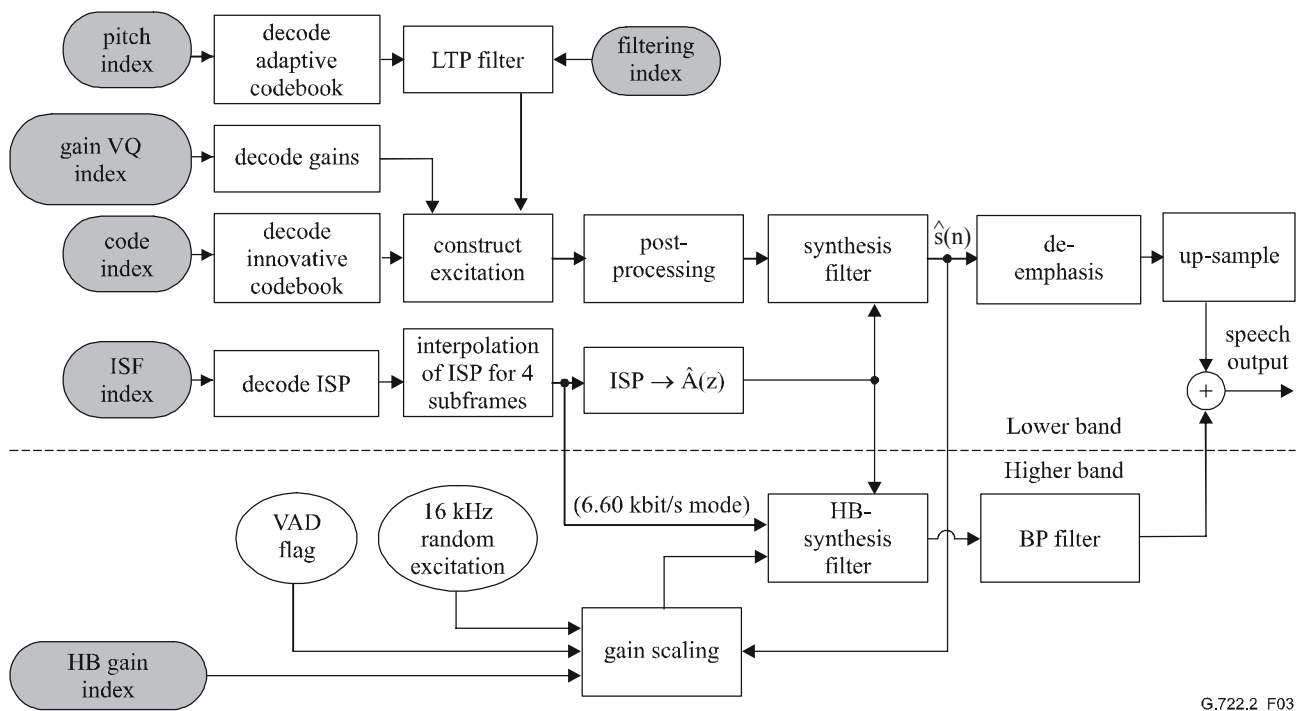


Figure 3/G.722.2 – Detailed block diagram of the ACELP decoder

4.5 Sequence and subjective importance of encoded parameters

The encoder will produce the output information in a unique sequence and format, and the decoder must receive the same information in the same way. In Tables 12a-12i, the sequence of output bits and the bit allocation for each parameter is shown.

The different parameters of the encoded speech and their individual bits have unequal importance with respect to subjective quality.

5 Functional description of the encoder

In this clause, the different functions of the encoder represented in Figure 2 are described.

5.1 Preprocessing

The encoder performs the analysis of the LPC, LTP and fixed codebook parameters at 12.8-kHz sampling rate. Therefore, the input signal has to be decimated from 16 kHz to 12.8 kHz. The decimation is performed by first upsampling by 4, then filtering the output through low-pass FIR filter $H_{\text{decim}}(z)$ that has the cut-off frequency at 6.4 kHz. Then, the signal is downsampled by 5. The filtering delay is compensated by adding zeros into the end of the input vector.

After the decimation, two preprocessing functions are applied to the signal prior to the encoding process: high-pass filtering and pre-emphasizing (and signal down-scaling).

(Down-scaling consists of dividing the input by a factor of 2 to reduce the possibility of overflows in the fixed-point implementation.)

The high-pass filter serves as a precaution against undesired low frequency components. A filter at a cut-off frequency of 50 Hz is used, and it is given by:

$$H_{h1}(z) = \frac{0.989502 - 1.979004z^{-1} + 0.989502z^{-2}}{1 - 1.978882z^{-1} + 0.9799126z^{-2}} \quad (4)$$

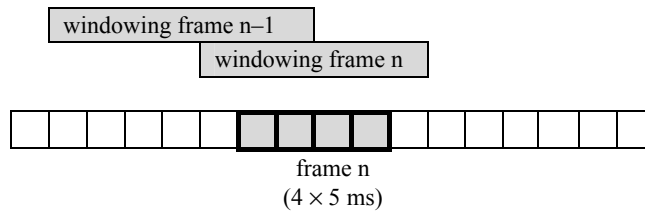
(Both down-scaling and high-pass filtering are combined by dividing the coefficients at the numerator of $H_{hl}(z)$ by 2.)

In the pre-emphasis, a first order high-pass filter is used to emphasize higher frequencies, and it is given by:

$$H_{\text{pre-emph}}(z) = 1 - 0.68z^{-1} \quad (5)$$

5.2 Linear prediction analysis and quantization

Short-term prediction, or LP, analysis is performed once per speech frame using the autocorrelation approach with 30-ms asymmetric windows. An overhead of 5 ms is used in the autocorrelation computation. The frame structure is depicted below.



The autocorrelations of windowed speech are converted to the LP coefficients using the Levinson-Durbin algorithm. Then the LP coefficients are transformed to the ISP domain for quantization and interpolation purposes. The interpolated quantized and unquantized filters are converted back to the LP filter coefficients (to construct the synthesis and weighting filters at each subframe).

5.2.1 Windowing and autocorrelation computation

LP analysis is performed once per frame using an asymmetric window. The window has its weight concentrated at the fourth subframe and it consists of two parts: the first part is a half of a Hamming window and the second part is a quarter of a Hamming-cosine function cycle. The window is given by:

$$\begin{aligned} w(n) &= 0.54 - 0.46 \cos\left(\frac{2\pi n}{2L_1 - 1}\right), & n = 0, \dots, L_1 - 1 \\ &= \cos\left(\frac{2\pi(n - L_1)}{4L_2 - 1}\right), & n = L_1, \dots, L_1 + L_2 - 1 \end{aligned} \quad (6)$$

where the values $L_1 = 256$ and $L_2 = 128$ are used.

The autocorrelations of the windowed speech $s'(n)$, $n = 0, \dots, 383$ are computed by:

$$r(k) = \sum_{n=k}^{383} s'(n)s'(n-k), \quad k = 0, \dots, 16 \quad (7)$$

and a 60-Hz bandwidth expansion is used by lag windowing the autocorrelations using the window [3]:

$$w_{\text{lag}}(i) = \exp\left[-\frac{1}{2}\left(\frac{2\pi f_0 i}{f_s}\right)^2\right], \quad i = 1, \dots, 16 \quad (8)$$

where $f_0 = 60$ Hz is the bandwidth expansion and $f_s = 12\,800$ Hz is the sampling frequency. Further, $r(0)$ is multiplied by the white noise correction factor 1.0001 which is equivalent to adding a noise floor at -40 dB.

5.2.2 Levinson-Durbin algorithm

The modified autocorrelations $r'(0)=1.0001r(0)$ and $r'(k)=r(k)w_{lag}(k), k=1,\dots,16$, are used to obtain the LP filter coefficients $a_k, k=1,\dots,16$ by solving the set of equations.

$$\sum_{k=1}^{16} a_k r'(|i-k|) = -r'(i), \quad i=1,\dots,16 \quad (9)$$

The set of equations in (9) is solved using the Levinson-Durbin algorithm [3]. This algorithm uses the following recursion:

$$E(0)=r'(0)$$

For $i=1$ to 16 do

$$k_i = -\left[r'(i) + \sum_{j=1}^{i-1} a_j^{(i-1)} r'(i-j)\right] / E(i-1)$$

$$a_i^{(i)} = k_i$$

For $j=1$ to $i-1$ do

$$a_j^{(i)} = a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}$$

$$E(i) = (1 - k_i^2) E(i-1)$$

The final solution is given as $a_j = a_j^{(16)}, j=1,\dots,16$.

The LP filter coefficients are converted to the ISP representation [5] for quantization and interpolation purposes. The conversions to the ISP domain and back to the LP filter domain are described in 5.2.3 and 5.2.4.

5.2.3 LP to ISP conversion

The LP filter coefficients $a_k, k=1,\dots,16$, are converted to the ISP representation for quantization and interpolation purposes. For a 16th-order LP filter, the ISPs are defined as the roots of the sum and difference polynomials:

$$f_1'(z) = A(z) + z^{-16} A(z^{-1}) \quad (10)$$

and:

$$f_2'(z) = A(z) - z^{-16} A(z^{-1}) \quad (11)$$

respectively. (The polynomials $f_1'(z)$ and $f_2'(z)$ are symmetric and antisymmetric, respectively.) It can be proven that all roots of these polynomials are on the unit circle and they alternate each other [6]. $f_2'(z)$ has two roots at $z=1$ ($\omega=0$) and $z=-1$ ($\omega=\pi$). To eliminate these two roots, we define the new polynomials:

$$f_1(z) = f_1'(z) \quad (12)$$

and:

$$f_2(z) = f_2'(z) / (1 - z^{-2}) \quad (13)$$

Polynomials $f_1(z)$ and $f_2(z)$ have 8 and 7 conjugate roots on the unit circle ($e^{\pm j\omega_i}$), respectively. Therefore, the polynomials can be written as:

$$F_1(z) = (1 + a[16]) \prod_{i=0,2,\dots,14} (1 - 2q_i z^{-1} + z^{-2}) \quad (14)$$