High-Definition Telephony over Heterogeneous Networks

Von der Fakultät für Elektrotechnik und Informationstechnik
der Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Ingenieur

Bernd Geiser

aus Mönchengladbach

Berichter: Universitätsprofessor Dr.-Ing. Peter Vary
Universitätsprofessor Dr.-Ing. Jens-Rainer Ohm

Tag der mündlichen Prüfung: 19. April 2012

AACHENER BEITRÄGE ZU DIGITALEN NACHRICHTENSYSTEMEN

Herausgeber:

Prof. Dr.-Ing. Peter Vary
Institut für Nachrichtengeräte und Datenverarbeitung
Rheinisch-Westfälische Technische Hochschule Aachen
Muffeter Weg 3a
52074 Aachen
Tel.: 0241-80 26 956
Fax.: 0241-80 22 186

Bibliografische Information der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar

1. Auflage Aachen:
Wissenschaftsverlag Mainz in Aachen
(Aachener Beiträge zu digitalen Nachrichtensystemen, Band 33)
ISSN 1437-6768
ISBN 3-86130-339-6

Wissenschaftsverlag Mainz
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 2 39 48 oder 02 41 / 87 34 34
Fax: 02 41 / 87 55 77
www.Verlag-Mainz.de

Herstellung: Druckerei Mainz GmbH,
Süsterfeldstr. 83, 52072 Aachen
Tel.: 02 41 / 87 34 34; Fax: 02 41 / 87 55 77
www.Druckservice-Aachen.de

Gedruckt auf chlorfrei gebleichtem Papier

"D 82 (Diss. RWTH Aachen University, 2012)"

Danksagung

Die vorliegende Arbeit entstand im Rahmen meiner Tätigkeit als wissenschaftlicher Angestellter am Institut für Nachrichtengeräte und Datenverarbeitung der Rheinisch-Westfälischen Technischen Hochschule Aachen. An dieser Stelle möchte ich mich bei allen bedanken, die meine Forschungstätigkeit – und damit auch die vorliegende Dissertation – ermöglicht, unterstützt und gefördert haben.

Mein besonders herzlicher Dank gilt Herrn Prof. Dr.-Ing. Peter Vary für seine uneingeschränkte Unterstützung meiner Tätigkeiten und für die Betreuung dieser Dissertation. Herrn Prof. Dr.-Ing. Jens-Rainer Ohm danke ich für die Übernahme des Koreferats. Auch den jetzigen und früheren Kolleginnen und Kollegen des Instituts gilt mein herzlicher Dank für die freundschaftliche Zusammenarbeit und die angenehme Arbeitsatmosphäre. Insbesondere möchte ich diejenigen Kolleginnen und Kollegen hervorheben, die durch fachliche Kooperation, Anregungen oder auch Korrekturlesen zur vorliegenden Arbeit beigetragen haben: Herr Dr.-Ing. Hauke Krüger, Herr Dr.-Ing. Heiner Löllmann, Herr Dr.-Ing. Peter Jax, Frau Dipl.-Ing. Birgit Schotsch, Herr Dipl.-Ing. Magnus Schäfer, sowie Herr Dipl.-Ing. Thomas Schlien. Ich möchte mich weiterhin bei den Studentinnen und Studenten bedanken, die mit ihren Studien-, Bachelor-, Master- und Diplomarbeiten wichtige Beiträge geleistet haben. Ein weiterer Dank gebührt den Kooperations- und Ansprechpartnern bei den Firmen Siemens und Huawei. Die sehr erfolgreiche Zusammenarbeit im Rahmen verschiedener Forschungsprojekte war für diese Arbeit von unschätzbarem Wert. Schließlich danke ich ganz besonders meinen Eltern, die mir das Studium der Elektrotechnik ermöglicht haben und mich stets mit vollen Kräften unterstützt haben. Liebe Sylvia, auch für Deine Unterstützung und Geduld ein ganz herzliches Dankeschön von mir!

Aachen, im Mai 2012 *Bernd Geiser*

Abstract

As of today, the lion's share of the worldwide (fixed and mobile) telephone connections is still restricted to audio frequencies below 4 kHz, leading to the familiar sound character of "telephone speech." For a clearly improved speech quality and a new "sensation of presence," however, audio frequencies up to 7 kHz (or even more) would be needed. As a matter of fact, the required costly and time-consuming modifications of the existing network equipment turned out to be a major obstacle for the introduction of (long existing) high quality speech and audio coding techniques in today's networks. Currently, even if both end-user terminals are suitably equipped, the telephone network can effectively preclude a high quality audio reproduction.

It can be expected that these network limitations will prevail for a long time. To account for this situation, in this thesis, concepts, methods and algorithms are investigated, evaluated and compared that facilitate a major audio quality upgrade of existing speech communication systems while maintaining backwards compatibility with the installed infrastructure.

The thesis makes major contributions to the following three principal scenarios:

Bandwidth Extension for Embedded Speech and Audio Coding
Two new bandwidth extension (BWE) algorithms have been developed in the context of recent ITU-T standardization projects for embedded speech and audio coding:

- *Time Domain Bandwidth Extension* (TDBWE) of speech signals towards the "wideband" frequency range (50 Hz – 7 kHz). This algorithm has been standardized as a part of the recent VoIP codec ITU-T Rec. G.729.1 which extends the widely deployed G.729 narrowband codec.

- *Transform Domain Bandwidth Extension* of audio signals towards the "super-wideband" frequency range (50 Hz – 14 kHz). This algorithm has been proposed for standardization in ITU-T where it proved to be the only competitor to fulfill all quality requirements.

Artificial Bandwidth Extension without Auxiliary Information
In this case, missing audio frequencies are estimated from the received, band-limited signal alone. The application of statistical estimation techniques to the new parameter sets (which have been developed for embedded coding) is investigated. A consistent quality improvement over the band-limited signals is obtained, but the speech quality does not reach the level of the embedded codec.

Bandwidth Extension with Steganographic Parameter Transmission
The availability of (even a small amount of) additional information can dramatically improve the quality of state-of-the-art speech codecs. For the common case that a modification of the digital bitstream format is not allowed, a new solution is proposed here: Data hiding techniques are used to deliver the BWE information to the receiving terminal without altering the standard bitstream format. The inaudibility of the hidden information is ensured by a joint source encoding and data hiding procedure. As a practically relevant application, this concept is applied to ACELP (Algebraic Code Excited Linear Prediction) codecs as used in GSM/UMTS mobile telephony. The key advantage of the proposed solution is its full backwards compatibility with the standard narrowband codecs, i.e., the existing network infrastructure can be kept and used without any modifications.

Contents

Notation and Symbols

In this thesis, the following conventions are used to denote quantities: Capital bold letters refer to matrices (e.g., $\mathbf{X}$), bold letters refer to vectors (e.g., $\mathbf{x}$), scalar values are not bold (e.g., x). Quantized or estimated variables are marked with a hat (e.g., $\hat{x}$). In contrast, a tilde (e.g., $\tilde{x}$) marks intermediate signals or, mainly in Chapter 5, signals with hidden data. Averaged values are denoted with a bar (e.g., $\bar{x}$). The frequency domain representation of a time domain (time-discrete) signal $s(k)$, with the discrete time index k, is written with the corresponding capital letter $S(\mu)$ for the μ-th frequency bin.

List of Principal Symbols

$a_i(\lambda)$	The i-th autoregressive coefficient in frame λ
$A(z)$	System function of a linear predictor
$\mathbf{b}$	ACELP pulse position likelihood vector
$\mathbf{c}_i$	CELP fixed codebook vector
$C^2(\mathbf{c})$	Numerator of the CELP criterion
$\mathcal{C}$	A (vector) codebook
$\mathcal{C}_m$	m-th sub-codebook for steganography
$\mathbf{d}$	CELP backward filtered target vector
$d(\cdot,\cdot)$	A distance measure between two vectors
$\dim(\mathbf{x})$	Dimension of the vector $\mathbf{x}$
$\mathrm{DFT}(\cdot)$	Discrete Fourier Transform of a time domain signal
$\delta(x)$	Kronecker symbol / unit pulse
Δ	Step size of a scalar quantizer
$E(\mathbf{c})$	Denominator of the CELP criterion
$f(\lambda)$	Binary flag for synthesis of spectral details
f_s	Sampling rate of a baseband or extension band signal
f'_s	Sampling rate of the input (or bandwidth extended) signal
f_c	Cutoff frequency
$F(\lambda, m)$	Logarithmic subband gain

$\varphi_{xx}(k)$	Autocorrelation function of the signal $x(k)$
$\mathbf{\Phi}$	CELP impulse response correlation matrix
$g(\lambda, \lambda_{\mathrm{SF}})$	Subframe gain
$g'(\lambda, \lambda_{\mathrm{SF}})$	Subframe gain of a received / temporally normalized signal
$g_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}})$	Relative correction gain factor for subframes
$g_{\mathrm{TGF}}(k)$	Temporal gain function
g_{a}	CELP adaptive codebook gain
g_{f}	CELP fixed codebook gain
$\gamma(\lambda, m)$	Subband gain
$\gamma'(\lambda, m)$	Subband gain of a received signal
$\gamma_{\mathrm{rel}}(\lambda, m)$	Relative correction gain factor for subbands
$\gamma_{\mathrm{SGF}}(\lambda, \mu)$	Spectral gain function
$\mathcal{G}(\cdot)$	Gray index assignment
$h(\cdot)$	Differential entropy
$h_0(k)$	Prototype low-pass filter impulse response
$h_{\mathrm{FBE}}(k)$	Impulse response of a filterbank equalizer
$h_{\mathrm{FBE}}^{(m)}(k)$	Impulse response of the m-th filterbank equalizer channel
$H(\cdot)$	Discrete entropy
$H(z)$	System function of a weighted LPC synthesis filter
$\mathbf{H}$	CELP weighted synthesis filter matrix
i_n	Index of the n-th pulse of an ACELP codebook
$I(\cdot, \cdot)$	Mutual information
$\jmath$	The imaginary unit
k	Discrete time index at sampling rate f_{s}
k'	Discrete time index at sampling rate f_{s}'
L	Frame length / frame shift
L_{FBE}	Filter length of a filterbank equalizer
L_{o}	Overlap length
L_{SF}	Subframe length / subframe shift
L_{w}	Window length
λ	Frame index
λ_{SF}	Subframe index
$\mathrm{ld}\,(\cdot)$	Logarithm with a basis of 2
$\log(\cdot)$	Logarithm with a basis of 10
m	Subband index

m (Chapter 5)	Steganographic message
M	Number of possible steganographic messages
M_s	Subband shift / spacing
M_SB	Width of a spectral subband
μ	Discrete frequency index (bin)
$\mathbb{M}$	Set of steganographic messages
$N_\mathcal{C}$	Number of codebook entries
N_L	Number of leafs of a graph tree
N_N	Number of nodes of a graph tree
N_P	Number of pulses in an ACELP track
N_SE	Number of spectral envelope parameters per frame
N_T	Number of tracks in an ACELP codec
N_TE	Number of temporal envelope parameters per frame
$\mathbb{N}$	The natural numbers
p_n	Pulse position of th n-th pulse in an ACELP codec
$p(\cdot)$	A probability density function
$p(\lambda)$	Pitch / harmonic grid parameter
$\mathbf{p}(\lambda)$	Parameter set/vector for bandwidth extension
$p_\mathrm{offset}(\lambda)$	Pitch / harmonic grid offset parameter
$P(\cdot)$	Probability of an event
$\mathcal{P}$	A subset of ACELP pulses
R_DH	Hidden data rate
$\mathbb{R}$	The real numbers
$s_\mathrm{p}(\cdot)$	Sign of an ACELP pulse
$s^\mathrm{t}(k)$	Temporally normalized signal
$s^\mathrm{w}(k)$	A windowed signal
$s_\mathrm{bb}(k)$	Baseband signal
$s_\mathrm{bwe}(k')$	Bandwidth extended signal
$s_\mathrm{eb}(k)$	Extension band signal
$s_\mathrm{hb}(k)$	High extension band signal (4 – 8 kHz)
$s_\mathrm{nb}(k)$	Narrowband signal
$s_\mathrm{swb}(k')$	Super-wideband signal
$s_\mathrm{uhb}(k)$	Upper high extension band signal (8 – 16 kHz)
$s_\mathrm{wb}(k')$	Wideband signal
$\mathrm{sign}(x)$	Sign of the value x

$S^{\mathrm{DFT}}(\lambda, \mu)$	Discrete Fourier Transform of frame λ of the signal $s(k)$
$S^{\mathrm{MDCT}}(\lambda, \mu)$	Modified Discrete Cosine Transform of frame λ of $s(k)$
$S^{\mathrm{PS}}(\lambda, \mu)$	MDCT Pseudo Spectrum of frame λ of the signal $s(k)$
σ_x^2	Variance of the signal x
t	An ACELP track index
$t(\lambda)$	Binary transient flag
$T(\lambda, \lambda_{\mathrm{SF}})$	Logarithmic subframe gain
$\mathcal{T}_t$	ACELP track
$\mathcal{T}_t^m$	Restricted ACELP track for hiding the message m
$\tau(\lambda)$	Inverse tonality parameter
$\mathbf{u}_j$	CELP adaptive codebook vector
$u(k)$	Excitation signal
$u^{\mathrm{LP}}(k)$	The linear prediction residual
$\mathbf{v}$	CELP target signal
$w_{\mathrm{F}}(k)$	Prototype filter / window function for frequency transforms
$w_{\mathrm{LP}}(k)$	Window function for LPC analysis
$w_{\mathrm{T}}(k)$	Time domain window function for interpolation
$w_{\mathrm{XF}}(k)$	Window function for crossfading
$W(\mu)$	Frequency domain window function
$W_{\mathrm{S}}(\mu)$	Frequency domain window function for interpolation
$W(z)$	System function of a perceptual weighting filter
$\hat{x}_{\mathrm{MMSE}}$	MMSE estimate of the value x
$\mathbf{x}_{\mathrm{f}}$	Feature vector
$\mathbf{X}_{\mathrm{f}}$	Feature vector sequence
$\mathbb{Z}$	The set of integer numbers

Glossary

Δ-PEAQ	Difference in ↗**PEAQ** scores
Δ-PESQ	Difference in ↗**PESQ** scores
Δ-PESQ$_{\mathbf{seg}}$	Difference in ↗**PESQ$_{\mathbf{seg}}$** scores
3GPP	3rd Generation Partnership Project
3GPP2	3rd Generation Partnership Project 2
AAC	↗**MPEG** Advanced Audio Codec
AAC-ELD	Enhanced Low Delay ↗**AAC**
ABWE	Artificial ↗**BWE**
ABX	ABX: A subjective listening test method
ACB	Adaptive Codebook
ACELP	Algebraic ↗**CELP**
ACF	Autocorrelation Function
ACR	Absolute Category Rating
ADPCM	Adaptive Differential ↗**PCM**
AMR	↗**3GPP** Adaptive Multirate Codec
AMR-WB	↗**3GPP** Adaptive Multirate Wideband Codec
AMR-WB+	↗**3GPP** Adaptive Multirate Wideband Plus Codec
AR	Autoregressive
ATE	Adaptive Temporal Envelope
Amd.	Amendment
bb	Baseband
BSDH	Bitstream Data Hiding
BWE	Bandwidth Extension
C	C Programming Language
CDF	Cumulative Density Function
CELP	Code Excited Linear Prediction
CELT	Constrained Energy Lapped Transform Codec
CRC	Cyclic Redundancy Check
CS-ACELP	Conjugate Structure ↗**ACELP**
CuT	Codec under Test

dBov	Decibel w.r.t. the overload point
DCR	Degradation Category Rating
DEMUX	Demultiplexer
DFT	Discrete Fourier Transform
DH	Data Hiding
DMOS	Degradation ↗**MOS**
DWM	Digital Watermarking
eb	Extension Band
EBU	European Broadcast Union
EFR	↗**3GPP** Enhanced Full Rate Codec
E^{2}FR	Enhanced ↗**EFR**
EPS	Evolved Packet System
eSBR	Enhanced ↗**SBR**
ETSI	European Telecommunication Standardization Institute
EVRC	↗**3GPP2** Enhanced Variable Rate Codec
EVRC-WB	↗**3GPP2** ↗**EVRC** Wideband Codec
FB	Full Band
FBE	Filterbank Equalizer
FCB	Fixed Codebook
FDM	Frequency Division Multiplex
FEC	Frame Erasure Concealment
FER	Frame Erasure Rate
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FR	↗**3GPP** Full Rate Codec
G.191	↗**ITU-T** Software Tools
G.711	↗**ITU-T** narrowband A/μ-law codec
G.718	Hierarchical ↗**ITU-T** wideband codec
G.719	Low-complexity ↗**ITU-T** full band codec
G.722	↗**ITU-T** ↗**ADPCM** wideband codec
G.722.1	Low-complexity ↗**ITU-T** wideband transform codec
G.722.1C	Low-complexity ↗**ITU-T** super-wideband transform codec
G.722.2	↗**ITU-T** codec identical to ↗**AMR-WB**
G.723.1	↗**ITU-T** low bit rate codec for multimedia communication
G.726	↗**ITU-T** ↗**ADPCM** codec

G.729	↗**ITU-T** ↗**CS-ACELP** codec
G.729.1	Embedded ↗**ITU-T** codec for ↗**VoIP**
G.729.1-SWB	Super-wideband extension of ↗**G.729.1**
G.729A	Low-complexity variant of ↗**G.729**
GDFT	Generalized ↗**DFT**
GLCVQ	Gosset Low Complexity ↗**VQ**
GMM	Gaussian Mixture Model
GSM	Global System for Mobile Communications
GSM 06.10	↗**FR**
GSM 06.60	↗**EFR**
GSM 06.90	↗**AMR**
H.264	↗**ITU-T** standard for video compression
hb	High band
HD	High Definition
HE-AAC	High Efficiency ↗**AAC**
HMM	Hidden Markov Model
HP	High Pass
HWR	Host-to-Watermark Ratio
IDFT	Inverse ↗**DFT**
IIR	Infinite Impulse Response
IMDCT	Inverse ↗**MDCT**
ISDN	Integrated Services Digital Network
ISPP	Interleaved Single Pulse Permutation
ITU	International Telecommunication Union
ITU-R	↗**ITU** Radiocommunication Sector
ITU-T	↗**ITU** Telecommunication Standardization Sector
JPEG	Joint Photographic Experts Group
JSCDH	Joint Source Coding and Data Hiding
KLT	Karhunen-Loève Transform
LBG	Linde-Buzo-Gray Algorithm
LP	Linear Prediction **or** Low Pass
LPC	Linear Predictive Coding
LSB	Least/Less Significant Bit(s)
LSD	Log-Spectral Distortion
LSF	Line Spectral Frequencies

LSP	Line Spectral Pairs
LTE	Long Term Evolution of ↗**UMTS**
LTP	Long Term Prediction
M-DMOS	Modified ↗**DMOS**
MAP	Maximum-a-Posteriori
MDCT	Modified Discrete Cosine Transform
MELP	Mixed-Excitation Linear Prediction
MFCC	Mel-Frequency Cepstral Coefficients
MMSE	Minimum ↗**MSE**
MOS	Mean Opinion Score
MOS-LQO	↗**MOS** - Listening Quality Objective
mp3PRO	MP3 audio codec with ↗**SBR** tools
MPEG	Moving Picture Experts Group
MSB	Most/More Significant Bit(s)
MSE	Mean Square Error
MUX	Multiplexer
NB	Narrowband
NGMN	Next Generation Mobile Networks
NGN	Next Generation Networks
NTT	Nippon Telegraph and Telephone
ODG	Objective Difference Grade
P.341	↗**ITU-T** wideband and super-wideband filters
P.48	↗**ITU-T** telephone band filters
P.56	↗**ITU-T** speech voltmeter
PCM	Pulse Code Modulation
PDF	Probability Density Function
PE	Phase Equalizer
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
$PESQ_{seg}$	segmental ↗**PESQ**
POTS	Plain Old Telephone System
PS	Pseudo Spectrum
QIM	Quantization Index Modulation
QMF	Quadrature Mirror Filter
RPE	Regular Pulse Excitation

RQ	Requantization
Rec.	Recommendation
Ref-A-B	Reference-A-B (subjective listening test method)
SAE	System Architecture Evolution
SBR	Spectral Band Replication
SE	Spectral Envelope
SF	Subframe
SG16	↗**ITU-T** Study Group 16
SGF	Spectral Gain Function
SNR	Signal-to-Noise Ratio
SQAM	Sound Quality Assessment Material
SWB	Super-Wideband
TB	Telephone Band
TCH/EFS	↗**GSM** enhanced full rate traffic channel
TDAC	Time Domain Alias Cancellation
TDBWE	Time Domain ↗**BWE**
TE	Temporal Envelope
TFO	Tandem-Free Operation
TGF	Temporal Gain Function
ToR	Terms of Reference
uhb	Upper high band
UMTS	Universal Mobile Telecommunications System
USAC	↗**MPEG** Unified Speech and Audio Coding
VMR-WB	↗**3GPP2** Variable-Rate Multimode Wideband codec
VQ	Vector Quantization
VoIP	Voice over Internet Protocol
WB	Wideband
WMOPS	Weighted Million Operations Per Second
WNR	Watermark-to-Noise Ratio
XF	Crossfading
ZCR	Zero Crossing Rate

Chapter 1

Introduction

These days, the telecommunication world is undergoing a major technology change which can be summarized by the catchphrase "Technology Convergence." This is usually understood as a universal network architecture for both fixed and mobile communications that will entirely rely on packet-switched data transmission. The main motivations behind the effort are presumably increased flexibility and cost-efficiency. But in particular for speech and audio communication applications, the opportunity should be seized to promote high quality services which are far superior to the long-accustomed speech telephony experience. However, it is already clear that large parts of the worldwide telephone network, to a large extent based on legacy technology, will prevail for a long time to come. This inevitably leads to a high degree of heterogeneity in network equipment as well as end user terminals. The present thesis investigates methods and algorithms that aim at a high quality speech and audio reproduction for the end user of such a heterogeneous communication network. Therefore, apart from signal parametrization, coding, and enhancement aspects, also the problem of backwards compatibility is addressed.

1.1 Evolution of Communication Networks

The invention and the commercialization of the telephone paved the road for the first global telecommunication network with all its infrastructure that has been constantly growing since. Over the years, the network has been enhanced with a myriad of different technologies and a vast variety of related standards. As a major step in this development, the network *digitalization* facilitated the deployment of new applications and services, cf. [Bellamy 1991]. Nowadays, an even more intense rise in new applications and services can be observed since the telephone world and the data communication world are increasingly converging. The consentient ultimate goal is the integration with the Internet which, being the second globally accessible communication network, has until recently evolved in parallel to the telephone network. As a consequence, telephony is rapidly changing from traditional *circuit switched* technology towards the more flexible and cost-efficient *packet switching* paradigm. In fact, telephony has already reached the Internet with the availability of *Voice over Internet Protocol* (VoIP) services, cf. [Hersent et al. 2005*b*] and [Hersent et al. 2005*a*].

The idea of a unified network architecture for both data transmission and real-time audio communication is also reflected by several projects and initiatives within the communications industry. Some keywords that are frequently encountered in this context are, e.g., *Next Generation Networks* (NGN), *Next Generation Mobile Networks* (NGMN), *fixed-mobile-convergence*, *triple play*, or an *All-IP* network. All of them address infrastructure and services for future communication networks. A concrete effort that is being pursued within the *3rd Generation Partnership Project* (3GPP) is the development and standardization of a so called *Evolved Packet System* (EPS), e.g., [Lescuyer & Lucidarme 2008] which accounts for both mobile access *and* the core network. The radio access infrastructure is being developed under the term LTE (*Long Term Evolution of UMTS*), e.g., [Sesia et al. 2009], while core network aspects are considered under the acronym SAE (*System Architecture Evolution*), e.g., [Olsson et al. 2009].

In the *backbone* of the current telephone network, most of the voice traffic is in fact already being routed over packet switched transmission links. However, the more peripheral parts of the network are mostly based on circuit switching and highly specialized technology, e.g., privately owned or government subnetworks as well as mobile radio access via GSM [Steele et al. 2001] or UMTS [Holma & Toskala 2004] speech channels. Support for all these network parts will have to be maintained while the new technology is being introduced.

1.2 Audio Quality in Speech Communication Systems

Over the last few years, data traffic has grown disproportionately compared to voice traffic, and—at least in mobile communications—this discrepancy is expected to increase. Nevertheless, voice communication will definitely retain its paramount social and economic importance and, compared to a typical data transmission scenario, special attention has to be paid to the service quality, in particular w.r.t. round-trip latency, availability, reliability, and audio quality.

A particularly important audio quality aspect which also represents the main focus of this thesis is the *reproduced audio bandwidth*. In addition, also the amount of local or remote acoustic noise, coding noise, and channel noise, as well as the suppression of acoustic echoes contribute to the overall audio quality. Moreover, the ability of the entire system to adapt to time-varying environmental conditions and transmission scenarios, i.e., its *flexibility*, plays a major role.

Telephone Speech

Despite the obvious relevance of improved speech and audio transmission, it can be observed that the quality and user experience of today's speech transmission systems is still far from optimum. In fact, during the past decades, the telephone audio quality could not be enhanced significantly, in particular w.r.t. the reproduced audio bandwidth.

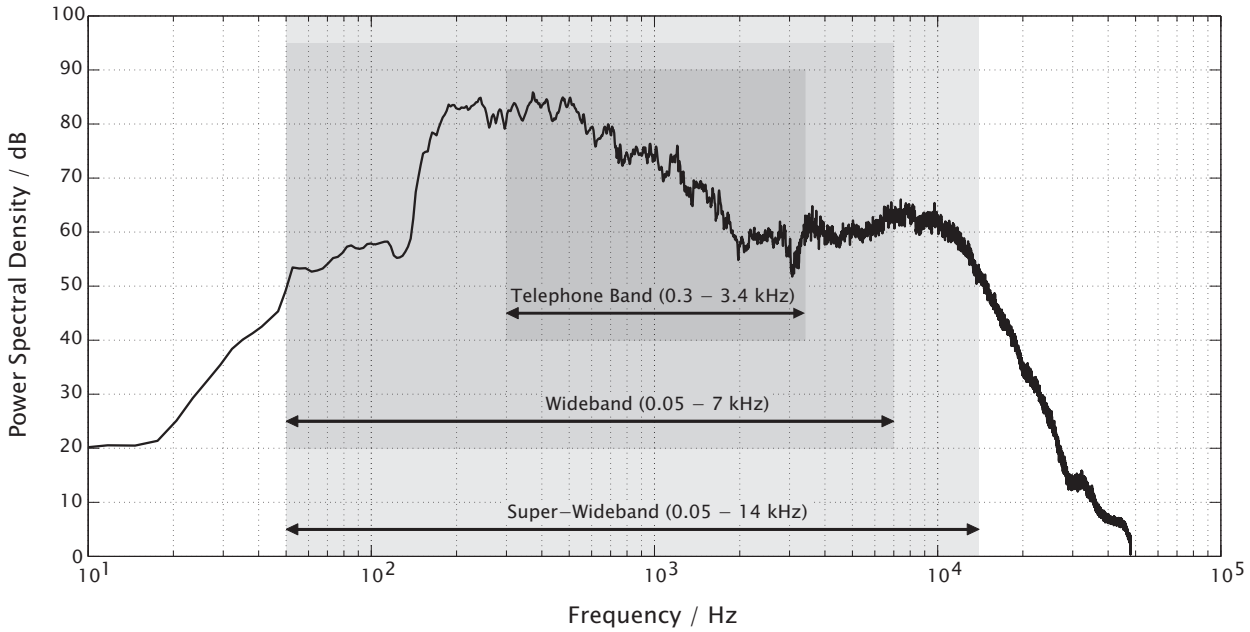


Figure 1.1: Bandwidth conventions for speech signals. The shown long-term power spectral density has been measured from 50 s of female speech, digitally recorded with a sampling rate of 96 kHz using a *Beyerdynamic MM1* measurement microphone.

Human speech is a spectrally rich signal covering (at least) the entire audible frequency range from a few Hertz up to approximately 20 kHz as illustrated in Figure 1.1. The most relevant part of this spectrum is located below 4 kHz and, for the basic requirement of intelligible natural language, "narrowband" speech within the traditional telephone frequency band from 300 Hz up to 3.4 kHz proved to be sufficient. Interestingly, with the first electrical speech transmissions over double-wire lines in the late 19th century, there was no strict limitation of the transmitted audio bandwidth. Yet, with higher line lengths (e.g., rural area loops), an increasing low-pass character of the transmitted speech had to be accepted. This problem was solved with the introduction of *loading coils*, i.e., discrete inductances placed at regular distances which effectively form a *filter* to equalize frequencies up to circa 3.5 kHz at the cost of strongly attenuated higher frequencies, cf. [Bellamy 1991, Figure 1.13]. Moreover, when *frequency division multiplexing* (FDM) techniques had to be introduced for the purpose of line sharing, a carrier spacing of 4 kHz was used for economical reasons and the transmitted bandwidth was henceforth strictly limited to less than 4 kHz. Later, with the introduction of *digital telephony*, the narrowband characteristic has been retained and consequently, the first *speech compression* standard for digital signals, ITU-T Rec. G.711 [ITU-T 1972], was designed to operate at a sampling rate of 8 kHz and a bit rate of 64 kbit/s. More advanced speech coding algorithms were then required for mobile telephony. For instance in the cellular GSM system, the bit rate could be lowered to 13 kbit/s [ETSI 1990, Vary et al. 1988] while maintaining an acceptable quality. Also here, the usual telephone bandwidth has been kept for compatibility reasons.

High Definition Telephony ...

The first attempt to provide a higher audio bandwidth and therefore a significantly improved quality to telephone customers has been initiated in 1984 with the standardization of the G.722 speech coding algorithm [ITU-T 1984]. This codec operates at the same bit rate as the old G.711 standard and provides so called "wideband" audio frequencies from 50 Hz up to 7 kHz, cf. Figure 1.1. Yet, G.722 was only intended for digital end-to-end connections and no dedicated network support was planned. Consequently, narrowband speech via G.711 remained state of the art. Similarly, for the case of mobile telephony over GSM and UMTS networks, it could be shown early that wideband speech coding is feasible at competitive bit rates [Paulus & Schnitzler 1996] and a suitable standard, the "Adaptive Multirate Wideband" speech codec (AMR-WB) [ETSI 2001*b*, Bessette et al. 2002], was finalized in 2001. However, the first careful endeavors to deploy this technology have only recently been made by network operators [Orange 2010].

Naturally, apart from mere audio bandwidth expansion, also a number of other quality aspects have to be taken into account to attain true "High Definition Telephony." Ultimately, multi-channel audio transmission appears desirable to facilitate *binaural* or *ambient audio communication* leading to a truly "immersive experience." However, since the main focus of this thesis is a wider reproduced audio bandwidth, most of these aspects are beyond scope and will only be referred to if required. Here, the term "High Definition Telephony" denotes a reproduced audio bandwidth that covers *at least* the wideband frequency range (50 Hz – 7 kHz). Yet, wideband speech is only the first step towards a "full band" audio transmission (typically 20 Hz – 20 kHz). A second intermediate step is shown in Figure 1.1. The so called "super-wideband" characteristic with its high frequency content from 7 kHz up to 14 kHz delivers additional clarity and a "sensation of presence." A number of codecs for this bandwidth has been developed recently, targeting applications with mixed speech and audio content, e.g., *high quality conferencing*, *e-learning*, *music on hold* or *remote monitoring*.

... over Heterogeneous Networks?

Apparently, high audio quality can already be achieved within *closed network environments*, as demonstrated by the commercially successful *Skype* VoIP software which transmits audio frequencies up to 12 kHz [Vos et al. 2010]. However, the corresponding large-scale modifications of the entire telephone network entail countless requirements and compatibility problems. Indeed, most of the industry initiatives listed in Section 1.1 do aim at "High Definition Telephony" for a *future* communication network, but little is done to improve the quality for *today's* network. Instead, "least common denominator" solutions are pursued, keeping up the status quo of narrowband speech. Although, at first sight, this might appear reasonable from the economic and marketing perspectives, it is nevertheless true that subscribers of new services will still experience inferior quality if their com-

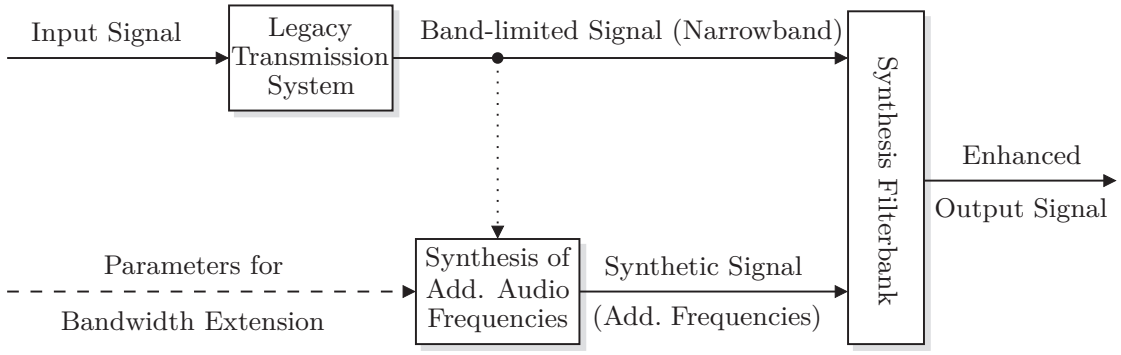


Figure 1.2: System for bandwidth extension (BWE) of band-limited speech or audio signals.

munication partner uses an old device or legacy network access. Hence, there is a definite need to account for the legacy parts of the network as well and, consequently, a highly heterogeneous network scenario has to be the basic assumption for all future developments.

Obviously, a heterogeneous transmission chain requires interoperability with legacy equipment. Therefore, any additional audio frequencies need to be supplied *outside* the legacy components. This demand can actually be fulfilled with techniques for parametric speech or audio *bandwidth extension* (BWE). The basic system setup for BWE is depicted in Figure 1.2. It is assumed that the input signal has been transmitted over a legacy link. It is therefore band-limited. The missing audio frequencies are then artificially regenerated based on a compact parametric description which may be obtained with or without the explicit transmission of auxiliary information. Finally, an enhanced output signal is produced with the help of a synthesis filterbank. In the following section, several transmission scenarios in a heterogeneous communication network will be introduced where bandwidth extension techniques can be successfully applied.

1.3 Application Scenarios for Bandwidth Extension

The main motivation for the application of bandwidth extension techniques can be seen in the desire that any new telephone (or communication device) with "High Definition Audio" support should be able to actually make use of its enhanced capabilities under all circumstances, i.e., even if the transmission chain involves legacy (narrowband) equipment or if the caller uses an older telephone without such support. Consequently, according to Figure 1.3, the following transmission scenarios are conceivable:

- **HD→HD→HD**
  An end-to-end "high definition" (HD) transmission chain is available that allows to transport a dedicated "high definition" speech or audio stream. This setup represents the desired, ideal scenario which, if ubiquitously available, marks the end of the ongoing technology change process.

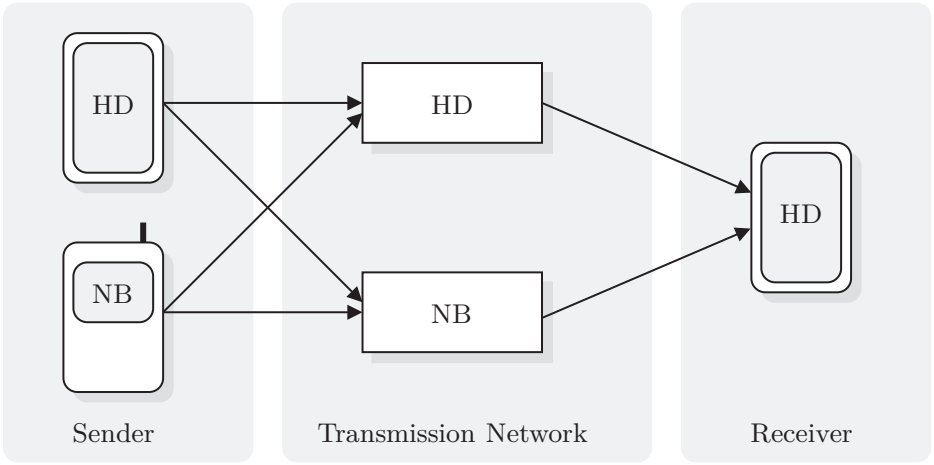


Figure 1.3: Transmission scenarios in a heterogeneous network (one-way).
HD: Device with HD Audio capability
NB: Device without HD Audio capability

- **HD→NB→HD**
  In this case, both terminals support HD transmission, but at least one intermediate component in the network is only suited for legacy narrowband telephony and inhibits the direct transmission of higher audio frequencies. Therefore, the bandwidth extension algorithm (Figure 1.2) has to be placed inside the receiving terminal or in a network component with HD audio support that is located near the end of the transmission chain.

  The parameters which are required for bandwidth extension have to be statistically estimated or, as a new proposal in this thesis, transmitted *inside* the legacy narrowband signal, which is accomplished with the help of *data hiding* techniques.

- **NB→HD→HD**
  Here, the calling terminal does not support HD transmission, but the network does. As no initial parameter set for bandwidth extension is available, statistical estimation techniques have to be applied. The respective algorithm can either be placed inside the receiving terminal or in a network component.

- **NB→NB→HD**
  Finally, it is possible that neither the network nor the sender support HD transmission. In this case, bandwidth extension and the related parameter estimation need to be incorporated in the receiving terminal.

The structure of the thesis, as outlined in the following, ensues from these application scenarios.

1.4 Thesis Outline

The main goal of this thesis is to provide a high quality speech and audio reproduction for the end users of heterogeneous transmission networks. The proposed methods and algorithms address typical use cases within the context of such networks. Thereby, specific signal parametrization, coding, enhancement, and transmission aspects as well as their applicability within the scenario of a heterogeneous network are discussed. The thesis is structured as follows.

Chapter 2: Signal Parametrization and Synthesis for Bandwidth Extension

To extend the reproduced audio bandwidth, first, a compact representation of the additional frequency content is required. Therefore, efficient signal parametrizations are introduced and suitable synthesis algorithms are devised. Both the extension towards wideband and super-wideband frequencies is considered.

Chapter 3: Bandwidth Extension for Embedded Speech and Audio Coding

As a first relevant use case for the techniques of Chapter 2, standardized and widely deployed codecs can be enhanced with a bandwidth extension algorithm by quantizing the respective parameter set and by appending additional "layers" to the bitstream. This approach is termed "embedded" or "hierarchical" coding, cf. [Geiser, Ragot & Taddei 2008, Erdmann 2005]. Two concrete algorithms that have been proposed in the context of international standardization projects are summarized, the first one targeting wideband speech transmission, the second one aiming at super-wideband speech *and* audio reproduction. Some of the described technologies have been incorporated into recent international standards for speech and audio coding.

Chapter 4: Receiver Based Parameter Estimation

In many scenarios, a quantized bandwidth extension parameter set is not available at the receiving terminal. Either the network may not have been able to transport the related bitstream layer or these parameters simply have not been determined in the sending terminal. In such cases, the quality of speech signals can be improved with a receiver-only modification where the parameter set is *estimated* based on the available information, i.e., from the narrowband speech signal alone. It is investigated to what extent parameter sets as described in Chapters 2 and 3 are amenable to a concise statistical estimation.

Chapter 5: Steganographic Parameter Transmission

Even a small amount of additional information can dramatically increase the possibilities to enhance the quality of older speech codecs. In this chapter, *data hiding techniques* are investigated to transmit this information to the receiving terminal,

while maintaining interoperability with existing network equipment and terminals. The inaudibility of the hidden information is ensured by a joint implementation of the data hiding procedure and the source coder. Apart from some basic conceptual considerations, a novel proposal for data hiding in state-of-the-art ACELP (Algebraic Code Excited Linear Prediction) codecs is devised and evaluated.

Chapter 6: Quality Evaluation and Comparison

The systems of Chapters 3 – 5 have been evaluated in a formal subjective listening test. The respective test results are discussed. To facilitate a meaningful comparison with existing telephony systems, standardized narrowband, wideband, and super-wideband codecs have been selected as reference conditions. In addition, results of instrumental quality measurements are provided for super-wideband bandwidth extension of *audio* signals.

Chapter 7: Summary

Finally, a summary and a closing discussion of the obtained results is given. Several practically relevant application scenarios are identified. The devised algorithms and techniques, for example, facilitate a major audio quality upgrade of current cellular networks.

Parts of the present thesis have been pre-published in the following references which I have authored or co-authored: [Geiser et al. 2006, Geiser et al. 2007a, Geiser & Vary 2007*a*, Geiser et al. 2007, Geiser & Vary 2007*b*, Geiser & Vary 2008*b*, Geiser, Mertz & Vary 2008, Geiser, Ragot & Taddei 2008, Geiser & Vary 2008*a*, Geiser et al. 2009, Geiser & Vary 2009, Geiser, Roggendorf & Vary 2010, Geiser, Krüger & Vary 2010, Geiser et al. 2011, Jax et al. 2006*a*, Jax et al. 2006*b*, Krüger et al. 2008, Krüger et al. 2010, Krüger et al. 2011*a*, Krüger et al. 2011*b*, Löllmann et al. 2009, Ragot et al. 2007, Thomas et al. 2010, Vary & Geiser 2007] — *B. Geiser*

Chapter 2

Signal Parametrization and Synthesis for Bandwidth Extension

This chapter focuses on efficient signal processing techniques to resynthesize the missing frequencies of band-limited speech or audio signals. As there are numerous alternative approaches to accomplish this goal, a selection of specific but practically relevant application scenarios is considered. The proposed algorithms are based on compact, parametric signal representations and constitute a common basis for the subsequent chapters of the present thesis. Some of the technologies that are described here have been incorporated into recent international standards for speech and audio coding. These specific realizations are described in Chapter 3.

2.1 System Overview

The basic system setup which is considered throughout the thesis is illustrated in the block diagram of Figure 2.1. Since the main goal is to provide a higher reproduced audio bandwidth for the end users of heterogeneous transmission networks, a standard ("legacy") transmission system is considered as the core component. *Additional audio frequencies* are then provided by *parametric* signal synthesis.

The digital input signal of the system in Figure 2.1 is denoted $s(k')$ with the sample index $k' \in \mathbb{Z}$ and the sampling period k'/f_{s}', where f_{s}' is the sampling rate. First, $s(k')$ is decomposed into a *baseband signal* $s_{\mathrm{bb}}(k)$ and an *extension band signal* $s_{\mathrm{eb}}(k)$ by means of a two-channel filterbank with decimation, i.e., the subband signals are in general processed with a (common) reduced sampling rate f_{s}. The corresponding sample index in the subsampled domain is $k \in \mathbb{Z}$.

The baseband signal $s_{\mathrm{bb}}(k)$ is transmitted via a "legacy" communication link involving a standard speech or audio codec. The extension band signal $s_{\mathrm{eb}}(k)$ is not directly transmitted. Instead, it is resynthesized at the receiver side. The employed synthesis algorithm produces an approximate version $\hat{s}_{\mathrm{eb}}(k)$ of the extension band signal based on a compact parameter set $\hat{\mathbf{p}}(\lambda)$ for each signal frame with index $\lambda \in \mathbb{Z}$. The corresponding *reference* parameter set $\mathbf{p}(\lambda)$ can be determined at the transmitter side based on the original extension band signal $s_{\mathrm{eb}}(k)$. The analysis window length is L_{w} samples, and the frame length (or frame shift) is L samples. The parameter set $\mathbf{p}(\lambda)$ is considered to be "compact" if its dimension, i.e., the number of parameters per frame, is much smaller than the frame shift L.

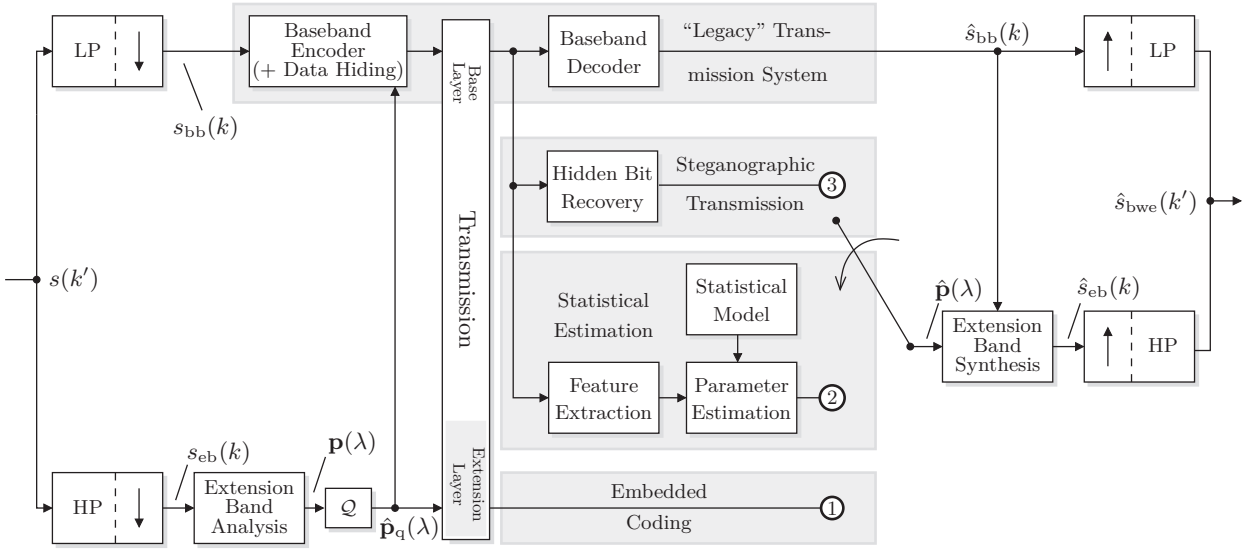


Figure 2.1: Overview of the considered transmission system. There are three possibilities to obtain the parameter set $\hat{\mathbf{p}}(\lambda)$:
(1) Embedded coding (Chapter 3)
(2) Statistical estimation (Chapter 4)
(3) Steganographic transmission (Chapter 5)

According to Figure 2.1, there are three ways to obtain $\hat{\mathbf{p}}(\lambda)$ at the receiver:

1. The parameter set $\mathbf{p}(\lambda)$ is *quantized* and the quantized representation $\hat{\mathbf{p}}_{\mathrm{q}}(\lambda)$ (or the related bits, respectively) are added to the baseband bitstream in the form of an *extension layer*. This approach, called "embedded coding" [Geiser, Ragot & Taddei 2008], is already pursued in a number of codec standards aiming at speech and also audio bandwidth extension. Two particular realizations are discussed in **Chapter 3**.

2. As another approach, the parameter set $\mathbf{p}(\lambda)$ can be *estimated* from the available baseband signal or from the parameters of the baseband codec with the help of a pre-trained statistical model. Meaningful models, however, can only be obtained for certain well-known source characteristics, i.e., the estimation approach is limited to *speech* signals. This is discussed in more detail in **Chapter 4**.

3. The third variant to obtain $\hat{\mathbf{p}}(\lambda)$ at the receiver, described in **Chapter 5**, makes use of *steganography*. The bits related to the quantized parameter set $\hat{\mathbf{p}}_{\mathrm{q}}(\lambda)$ are *hidden* in the bitstream of the narrowband codec. The decoder can recover these bits and supply a decoded parameter set $\hat{\mathbf{p}}(\lambda)$ to the extension band synthesizer.

In all three cases, the (received or estimated) parameter set $\hat{\mathbf{p}}(\lambda)$ is used to synthesize the signal $\hat{s}_{\mathrm{eb}}(k)$ which is then recombined with the received baseband signal $\hat{s}_{\mathrm{bb}}(k)$ to form the bandwidth extended output signal $\hat{s}_{\mathrm{bwe}}(k')$.

Table 2.1: Definition of frequency bands — (↓): downsampling applied

Name	Acronym	Lower cutoff [kHz]	Upper cutoff [kHz]	Sampling rate [kHz]
Full Band	FB	0.02	20	44.1 / 48
Super-Wideband	SWB	0.05	14	32
Wideband	WB	0.05	7	16
Narrowband	NB	0.05	4	8
Telephone Band	TB	0.3	3.4	8
WB Ext. Band	HB	4	7 / 8	8 (↓)
SWB Ext. Band	UHB	8	14	16 (↓)

To establish a common basis for the following chapters, this chapter focuses on the determination of the *reference parameters* $\mathbf{p}(\lambda)$ and on the corresponding *signal synthesis algorithms* to obtain the bandwidth extended output signal $\hat{s}_{\mathrm{bwe}}(k')$. Therefore, in the present chapter, the *unquantized* parameter set $\mathbf{p}(\lambda)$ is directly used at the receiver to produce $\hat{s}_{\mathrm{eb}}(k)$.

2.1.1 Baseband and Extension Band Definitions

The frequency bands that are used throughout this thesis are defined in Table 2.1. They mostly follow the definitions of the International Telecommunication Union (ITU). Note that the "narrowband" frequency range is defined as 0.05 – 4 kHz as opposed to the *telephone band* of 0.3 – 3.4 kHz. If possible, the filter characteristics that are recommended by ITU-T (P.48 [ITU-T 1976] and P.341 [ITU-T 1995]) are used to produce band-limited versions of the reference signals.

Two concrete scenarios for bandwidth extension are considered in this thesis:

- **Extension of NB/TB speech to WB speech**
  This scenario has, e.g., been studied in [Carl & Heute 1994, Jax 2002] for the case of *artificial* bandwidth extension (without auxiliary information). Here, the extension band signal covers audio frequencies from 4 kHz to 7 kHz and the baseband signal (in the NB case) provides the frequencies up to 4 kHz. If the baseband signal is limited by the telephone characteristic (TB) according to ITU-T Rec. P.48 [ITU-T 1976], effectively, a spectral gap from 3.4 kHz to 4 kHz remains. However, such a gap has no significant impact on the speech quality, see [Jax & Vary 2003, Pulakka et al. 2008]. Nevertheless, the missing low end of the speech (50 Hz – 300 Hz) has to be accounted for since it contributes significantly to the perceived naturalness of the speech signal. Some BWE methods that address a *low frequency regeneration* are referenced in Section 4.6.

- **Extension of WB audio signals to SWB audio signals**
  Bandwidth extension of *audio* signals has, e.g., been studied in [Dietz et al. 2002]. The extension band for the SWB case ranges from 8 kHz to 14 kHz. In the baseband, a wideband characteristic is enforced with the ITU-T Rec. P.341 filter [ITU-T 1995] with its passband from 50 Hz – 7 kHz. Consequently, a spectral gap between 7 kHz and 8 kHz remains. Again, for speech signals, there is only a negligible quality impact. However, for a concise reproduction of generic audio material (e.g., music), a dedicated module is required to fill the gap between 7 kHz and 8 kHz, e.g., [Geiser et al. 2009, Eksler & Jelínek 2011]. Here, this is not discussed for brevity.

Note that the step from SWB towards FB audio signals is not considered. However, similar signal processing techniques as used for the WB to SWB extension are expected to be appropriate for this task.

2.1.2 Filterbanks for Subband Analysis and Synthesis

In Figure 2.1, a digital *analysis filterbank* is used to split the input signal $s(k')$ into its baseband and extension band components. Likewise, a *synthesis filterbank* is required to recombine the synthesized extension band signal with the received baseband signal.

The use cases outlined in Section 2.1.1 require *two-channel* filterbanks with a band split frequency of $f_s/2$. Usually, *quadrature mirror filterbanks* (QMF-banks) are used for such purposes because of their perfect reconstruction properties, e.g., [Esteban & Galand 1977]. In the following, the Infinite Impulse Response (IIR) QMF-bank of [Löllmann & Vary 2008, Löllmann et al. 2009] with an allpass-based polyphase implementation as depicted in Figure 2.2(a) is used. Based on the polyphase allpass filters $A_0(z)$ and $A_1(z)$, the effective transfer function for the low and high pass analysis filters can be expressed as:

$$H_{\mathrm{LP}}(z) = A_0(z^2) + z^{-1}\,A_1(z^2) \tag{2.1}$$

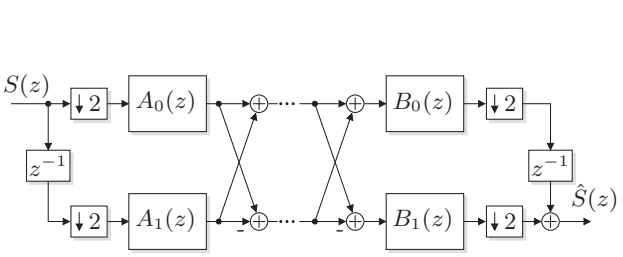
$$H_{\mathrm{HP}}(z) = A_0(z^2) - z^{-1}\,A_1(z^2). \tag{2.2}$$

The corresponding filter responses are shown in Figure 2.2(b). Similarly, the transfer functions of the effective synthesis filters can be written as

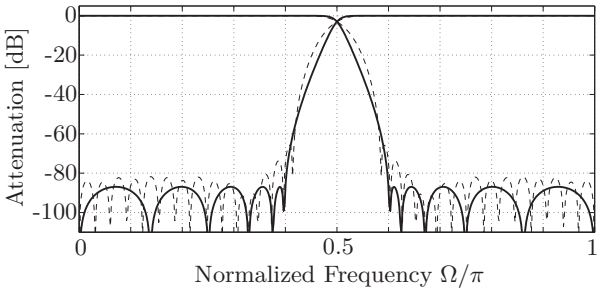
$$G_{\mathrm{LP}}(z) = z^{-1}\,B_0(z^2) - B_1(z^2) \tag{2.3}$$

$$G_{\mathrm{HP}}(z) = z^{-1}\,B_0(z^2) - B_1(z^2) \tag{2.4}$$

based on the polyphase filters $B_0(z)$ and $B_1(z)$. The synthesis filters are designed to cancel out the aliasing distortion so that a *perfect signal reconstruction* can be achieved if no spectral processing is applied. However, for the present IIR QMF-bank, only a "near-perfect" reconstruction can be realized because of its non-linear phase response. Therefore, the synthesis polyphase filters include an additional



(a) Polyphase network implementation of the allpass-based, two-channel IIR QMF-bank of [Löllmann et al. 2009].



(b) Magnitude responses of the analysis subband filters. Solid lines: IIR subband filters, dashed lines: ITU-T Rec. G.729.1 FIR reference.

Figure 2.2: IIR QMF-bank for subband analysis and synthesis.

phase equalizer term, see [Löllmann et al. 2009], which is designed to compensate the *subjectively objectionable* phase distortions. On the other hand, the IIR QMF-bank offers several advantages compared to a conventional finite impulse response (FIR) solution: a lower signal delay, lower complexity, and a slightly better filter characteristic (magnitude response, see Figure 2.2(b)).

It is worth noting that, for the present application, the QMF alias cancellation mechanism and hence the (near) perfect reconstruction properties are no longer effective. This can be explained by the fact that a *parametric signal regeneration*, as pursued in bandwidth extension algorithms, is a highly nonlinear operation. Instead, sufficiently steep filter characteristics with high stopband attenuation are more important than the strict adherence to the perfect reconstruction paradigm. Yet, (near) perfect reconstruction QMF-banks are still relevant for hierarchical audio codecs where an initial parametric signal is successively refined in terms of signal-to-noise ratio (SNR), cf. Chapter 3.

2.1.3 Analysis and Synthesis of the Extension Band Signal

Parametric signal synthesis for bandwidth extension is motivated by the insensitivity of the *human auditory system* towards phase distortions and a (limited) mismatch of *spectral details* at higher audio frequencies. Therefore, it is usually sufficient to reproduce a rough approximation of the spectral details, cf. the discussion in Section 2.5. However, certain coarse signal characteristics, in particular *spectral and temporal envelopes* need to be accurately represented to obtain a high subjective quality. Previous work on the topic often concentrated on *spectral* envelope parameters, e.g., [Jax 2002]. In this thesis, an additional, *explicit* description of the *temporal* envelope is proposed.

Figure 2.3 shows signal processing architectures that are suitable for parametric analysis and synthesis in the context of audio bandwidth extension. For the analysis of $s_{\mathrm{eb}}(k)$, i.e., to obtain the reference parameter vector $\mathbf{p}(\lambda)$, two algorithmic configurations are of interest, namely serial and parallel analysis as depicted

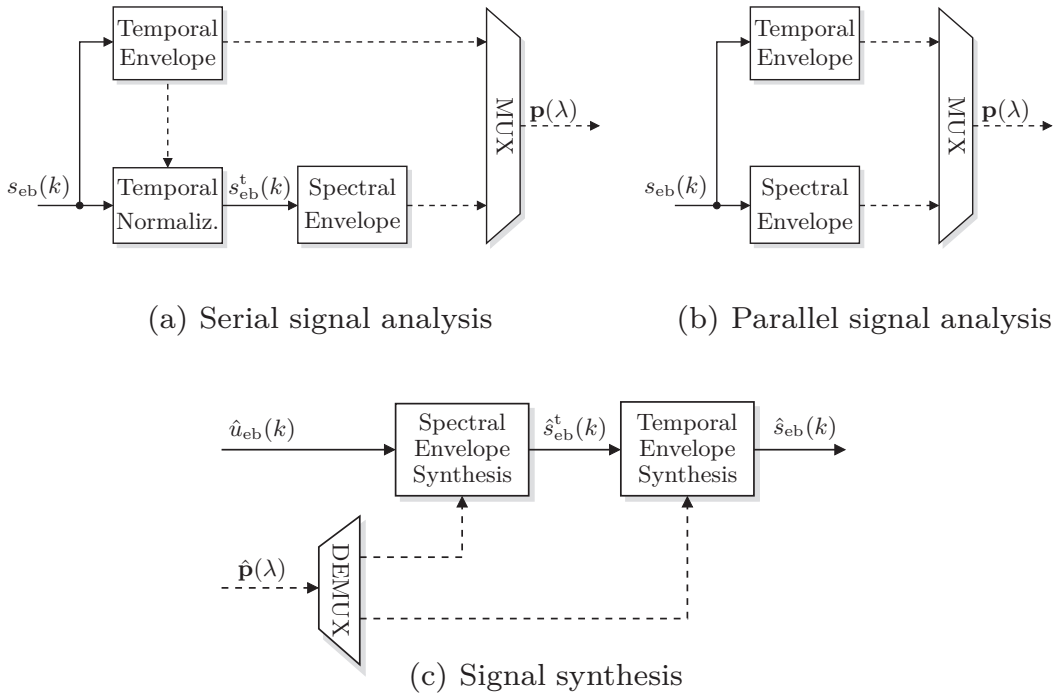


Figure 2.3: Analysis and synthesis of temporal and spectral envelopes.

in Figure 2.3(a) and Figure 2.3(b), respectively. Thereby, the serial analysis approach (with temporal normalization) is particularly interesting for an application in audio codecs because the temporally normalized signal $s^{\mathrm{t}}_{\mathrm{eb}}(k)$ has a reduced dynamic range and is therefore easier to encode, see Section 2.2. The corresponding synthesis algorithm is shown in Figure 2.3(c). The "excitation signal" $\hat{u}_{\mathrm{eb}}(k)$ is first spectrally shaped according to a spectral envelope description. Afterwards, the temporal envelope of $\hat{s}^{\mathrm{t}}_{\mathrm{eb}}(k)$ is restored. If the parallel analysis structure of Figure 2.3(b) is used, it is also possible to interchange the order of temporal and spectral signal shaping, see Section 3.2 for a corresponding application.

The parameters to describe the *temporal envelope* of $s_{\mathrm{eb}}(k)$ are introduced in Section 2.2. For the *spectral envelope*, two different parametrization approaches are investigated. The first (Section 2.3) is based on time-domain modeling while the second (Section 2.4) operates in a transformed domain. For these parameters, appropriate signal synthesis modules are devised. Also, the amount of *detail* concerning temporal and spectral envelopes that is required for speech (and audio) signals is investigated.

2.2 Temporal Envelope Representation and Control

A concise reproduction of the *gain contour* in the extension band is essential for a high quality bandwidth extension because high audio frequencies are particularly susceptible to temporal artifacts, e.g., [Taori et al. 2000, Kim et al. 2007]. Therefore, an explicit representation of the temporal envelope is proposed here and included in the parameter set $\mathbf{p}(\lambda)$.

2.2.1 Gain Function for Temporal (De-)Normalization

A simple, yet flexible solution to describe the temporal envelope of $s_{\mathrm{eb}}(k)$ are *subframe gains* $g(\lambda, \lambda_{\mathrm{SF}})$ which are for example used in [Geiser et al. 2007a] and [Geiser et al. 2009], see also Section 3.2 and Section 3.3. Therefore, each signal frame with index λ is subdivided in to N_{TE} subframes. The gains for these subframes with indices $\lambda_{\mathrm{SF}} \in \{0, \ldots, N_{\mathrm{TE}} - 1\}$ are then defined as

$$g(\lambda, \lambda_{\mathrm{SF}}) = \max \left\{ g_0, \sqrt{\frac{1}{L_{\mathrm{SF}}} \cdot \sum_{k=0}^{L_{\mathrm{SF}}-1} s_{\mathrm{eb}}^2(\lambda L + \lambda_{\mathrm{SF}} L_{\mathrm{SF}} + k)} \right\} \tag{2.5}$$

where L_{SF} is the subframe length which is required to divide the frame length L and $g_0 > 0$ is a fixed minimum gain value. The subframe gains $g(\lambda, \lambda_{\mathrm{SF}})$ (or their quantized counterparts $\hat{g}(\lambda, \lambda_{\mathrm{SF}})$) can be easily interpolated to form a "temporal gain function" (TGF) $g_{\mathrm{TGF}}(k)$ (or $\hat{g}_{\mathrm{TGF}}(k)$ in quantized form). A particularly efficient realization of the interpolation filter is an overlap-add using the slope of scaled Hann windows $w_{\mathrm{T}}(k) = [\sin(\pi(k+1)/(2\,L_{\mathrm{SF}}+2))]^2$:

$$\begin{aligned} &g_{\mathrm{TGF}}(\lambda L + \lambda_{\mathrm{SF}} L_{\mathrm{SF}} + k) \\ &\quad = \begin{cases} w_{\mathrm{T}}(k + \frac{L_{\mathrm{SF}}}{2}) \cdot g(\lambda, \lambda_{\mathrm{SF}}) + \left[1 - w_{\mathrm{T}}(k + \frac{L_{\mathrm{SF}}}{2})\right] \cdot g(\lambda, \lambda_{\mathrm{SF}} - 1) \\ \qquad\qquad \text{if} \quad k \in \{0, \ldots, L_{\mathrm{SF}}/2 - 1\} \\ [1 - w_{\mathrm{T}}(k)] \cdot g(\lambda, \lambda_{\mathrm{SF}}) \;\; + w_{\mathrm{T}}(k) \cdot g(\lambda, \lambda_{\mathrm{SF}} + 1) \\ \qquad\qquad \text{if} \quad k \in \{L_{\mathrm{SF}}/2, \ldots, L_{\mathrm{SF}} - 1\} \end{cases} \end{aligned} \tag{2.6}$$

for $k \in \{0, \ldots, L_{\mathrm{SF}} - 1\}$ where, for convenience, $g(\lambda, -1) \doteq g_{\mathrm{TGF}}(\lambda - 1, N_{\mathrm{TE}} - 1)$ and $g(\lambda, N_{\mathrm{TE}}) \doteq g_{\mathrm{TGF}}(\lambda + 1, 0)$. To compute (2.6), only one subframe of look-ahead is required. In the *serial* signal analysis approach according to Figure 2.3(a), the (quantized) TGF $\hat{g}_{\mathrm{TGF}}(k)$ is used to normalize the input signal $s_{\mathrm{eb}}(k)$:

$$s_{\mathrm{eb}}^{\mathrm{t}}(k) = s_{\mathrm{eb}}(k) \cdot \hat{g}_{\mathrm{TGF}}^{-1}(k). \tag{2.7}$$

While this multiplication considerably reduces the signal dynamics, an undesired side effect is that the spectrum components of the input signal are modified by a cyclic convolution with the Fourier transform of the inverse gain function (spectral leakage). To limit the impact on the spectrum components to the lowest possible amount, the interpolation window $w_{\mathrm{T}}(k)$ is designed such that $g_{\mathrm{TGF}}(k)$ exhibits sufficient low-pass characteristics, which is explained below, see Figure 2.4(c).

In the *synthesis* module of Figure 2.3(c), the signal $\hat{s}_{\mathrm{eb}}^{\mathrm{t}}(k)$ is analyzed in the same manner as specified by (2.5). The derived subframe gains $g'(\lambda, \lambda_{\mathrm{SF}})$ describe the

observed temporal envelope of $\hat{s}^{\mathrm{t}}_{\mathrm{eb}}(k)$. Then, together with the received (or estimated) parameters $\hat{g}(\lambda, \lambda_{\mathrm{SF}})$, relative gain factors

$$\hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}}) = \frac{\hat{g}(\lambda, \lambda_{\mathrm{SF}})}{g'(\lambda, \lambda_{\mathrm{SF}})} \tag{2.8}$$

can be determined. Following (2.6), the gains $\hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}})$ are used to construct the gain function $\hat{g}'_{\mathrm{TGF}}(k)$ to correct the temporal envelope of the signal $\hat{s}^{\mathrm{t}}_{\mathrm{eb}}(k)$:

$$\hat{s}_{\mathrm{eb}}(k) = \hat{s}^{\mathrm{t}}_{\mathrm{eb}}(k) \cdot \hat{g}'_{\mathrm{TGF}}(k). \tag{2.9}$$

Again, through interpolation by (2.6), it is ensured that $\hat{g}'_{\mathrm{TGF}}(k)$ exhibits a pronounced low-pass characteristic so that the impact on the spectrum of $\hat{s}^{\mathrm{t}}_{\mathrm{eb}}(k)$ remains tolerable. In the context of audio coding, (2.9) can be used to effectively suppress *pre-echo* artifacts that frequently occur with transform based codecs. This has been exploited in the concrete codec design that is described in Section 3.3.

Evaluation and Example

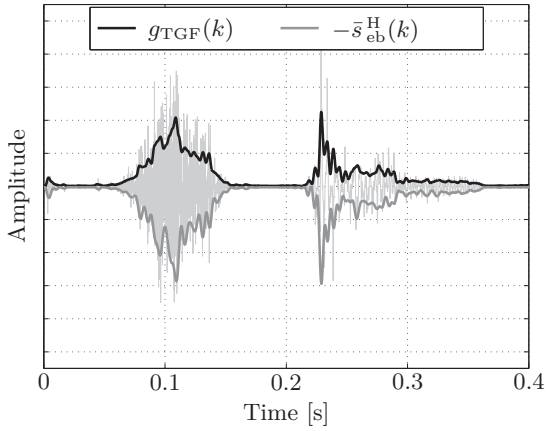
An example of the described temporal gain control mechanism is given in Figure 2.4 for a short signal segment $s_{\mathrm{eb}}(k)$ with $f_{\mathrm{s}} = 8\,\mathrm{kHz}$ and $L_{\mathrm{SF}} = 20 \mathrel{\hat{=}} 2.5\,\mathrm{ms}$.

In Figures 2.4(a) and 2.4(b) the proposed TGF approach is, for reference, compared with a commonly used alternative envelope representation, i.e., the *Hilbert envelope*

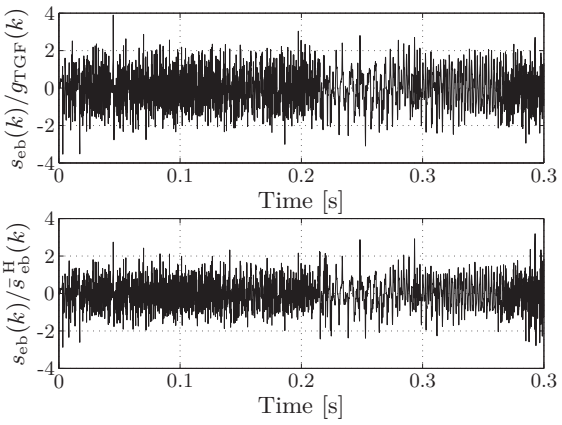
$$s^{\mathrm{H}}_{\mathrm{eb}}(k) = \left| s_{\mathrm{eb}}(k) + \jmath \cdot \sum_{i=-\infty}^{\infty} s_{\mathrm{eb}}(k-i) \frac{1-(-1)^i}{4\pi i} \right|. \tag{2.10}$$

In practice, the second summand of (2.10) needs to be approximated, e.g., using fast convolution in the DFT domain. The Hilbert envelope, which is also considered in [Kim et al. 2007], provides an estimate of the "instantaneous amplitude" of a signal, e.g., [Ohm & Lüke 2010]. Here, to facilitate a fair comparison with the TGF envelope of (2.6), a *low-pass filtered version* of $s^{\mathrm{H}}_{\mathrm{eb}}(k)$, denoted $\bar{s}^{\,\mathrm{H}}_{\mathrm{eb}}(k)$, is used. The cutoff frequency of the applied low-pass filter has been matched with the 6 dB cutoff frequency of the TGF interpolation window $w_{\mathrm{T}}(k)$, i.e., $f_{\mathrm{c}} = {}^{f_{\mathrm{s}}}/_{L_{\mathrm{SF}}}$. In Figure 2.4(a), the low-pass filtered Hilbert envelope is shown with a *negative sign*. Both envelope contours (i.e., $g_{\mathrm{TGF}}(k)$ and $\bar{s}^{\,\mathrm{H}}_{\mathrm{eb}}(k)$) are apparently quite similar (except that the maxima in $\bar{s}^{\,\mathrm{H}}_{\mathrm{eb}}(k)$ are more pronounced) and the normalization according to (2.7) actually results in a sufficiently flat signal $s^{\mathrm{t}}_{\mathrm{eb}}(k)$ for *both* approaches. This is shown in Figure 2.4(b). However, the TGF representation offers a number of practical advantages over the Hilbert envelope such as lower complexity and a lower algorithmic delay.

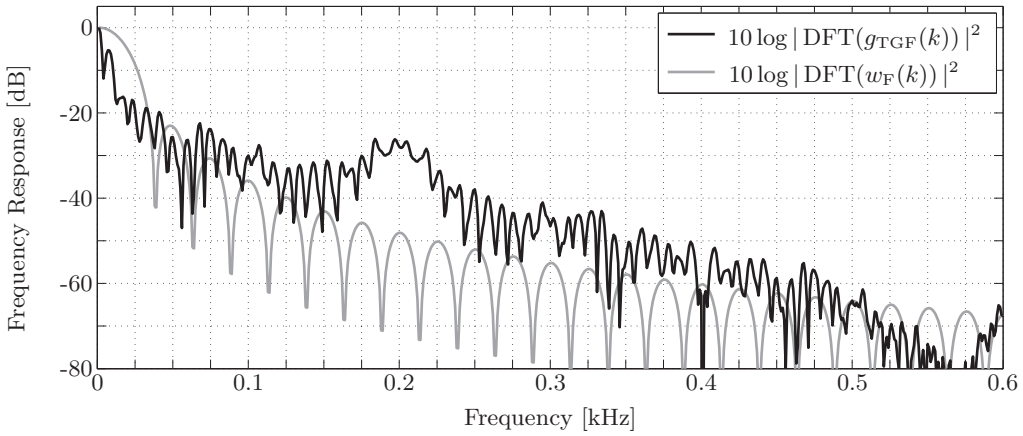
To limit the impact of the temporal normalization operation on the *spectral* components of the signal $s^{\mathrm{t}}_{\mathrm{eb}}(k)$, the gain function $g_{\mathrm{TGF}}(k)$ must exhibit sufficient low-pass characteristics. This is analyzed in Figure 2.4(c) based on the example signal segment of Figure 2.4(a). For reference, the frequency response of a



(a) TGF and filtered Hilbert envelope.



(b) Normalized signals.



(c) Frequency response of $g_{\mathrm{TGF}}(k)$ compared with the MDCT analysis window $w_{\mathrm{F}}(k)$ from ITU-T Rec. G.729.1 [ITU-T 2006, Ragot et al. 2007]. The channel spacing of the G.729.1 MDCT analysis is 25 Hz.

Figure 2.4: Example for temporal envelope modeling and temporal normalization ($f_{\mathrm{s}} = 8\,\mathrm{kHz}$, $L_{\mathrm{SF}} = 20 \triangleq 2.5\,\mathrm{ms}$).

standardized spectral analysis window $w_{\mathrm{F}}(k)$ is shown. This window is used for transform audio coding in the Modified Discrete Cosine Transform (MDCT) of ITU-T Rec. G.729.1. Here, the sampling frequency is also $f_{\mathrm{s}} = 8\,\mathrm{kHz}$.

Although the stopband attenuation of the window $w_{\mathrm{F}}(k)$ is somewhat better for the given parameter setting ($L_{\mathrm{SF}} = 20$), it can be concluded from Figure 2.4(c), that the spectral leakage of the temporal gain function $g_{\mathrm{TGF}}(k)$ remains well below the resolution of the spectral transform (which is actually used for transform audio coding). Therefore, despite the proposed temporal normalization, consistent results can be expected from a spectral analysis (e.g., with an MDCT) of the *temporally normalized signal* $s^{\mathrm{t}}_{\mathrm{eb}}(k)$. Nevertheless, it should be noted that the spectral analysis techniques to be described in Sections 2.3 and 2.4 are even less demanding since they only aim to describe the spectral *envelope*.

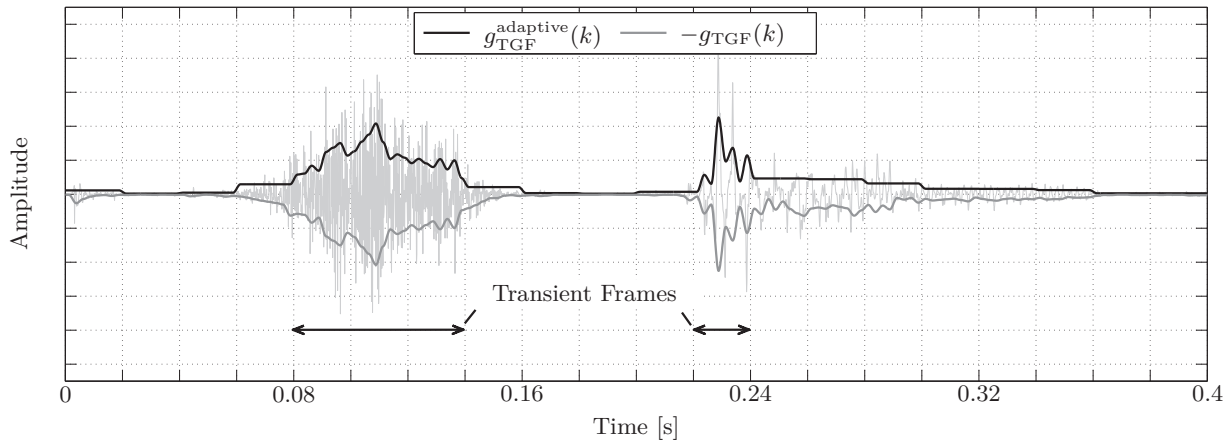


Figure 2.5: Temporal envelope modeling with signal-adaptive resolution ($f_\mathrm{s} = 8\,\mathrm{kHz}$, $L = 160 \mathrel{\hat{=}} 20\,\mathrm{ms}$, $L_\mathrm{SF} = 20 \mathrel{\hat{=}} 2.5\,\mathrm{ms}$, $N_\mathrm{TE} = 8$).

2.2.2 Temporal Envelope with Adaptive Resolution

To accommodate audio signals with strongly varying temporal characteristics, the temporal envelope resolution can be adapted to the current input characteristics by changing the subframe length L_SF and therefore the number N_TE of subframe gains per frame λ. Generally, two goals are pursued with this. First, the average number of parameters to describe the temporal envelope can be reduced significantly in contrast to permanently using the maximum temporal resolution, thus reducing the required bit rate. Second, an adapted temporal resolution avoids remaining (and unnecessary) spectral leakage effects.

As a practical and computationally efficient realization, *two* modes of operation can be employed. In this case, each audio frame of length L is classified as either "stationary" or "transient." Therefore, a simple yet effective transient detector can be used which determines if the maximum rising and/or falling slopes within the subframe gains $g(\lambda, \lambda_\mathrm{SF})$ of a frame λ exceed certain pre-specified thresholds. The temporal characteristic of *stationary* frames is then described by a *single gain factor* $g^{\mathrm{stat.}}(\lambda)$, i.e., $L_\mathrm{SF}^{\mathrm{stat.}} = L$ in (2.5). *Transient* segments require the *original* (higher) temporal resolution. It is important to note that the interpolation formula (2.6) can still be applied for stationary frames if the subframe gain parameters are held constant for the entire frame, i.e., $g(\lambda, \lambda_\mathrm{SF}) \equiv g^{\mathrm{stat.}}(\lambda)$. A typical example for the resulting *signal-adaptive* temporal gain function $g_\mathrm{TGF}^{\mathrm{adaptive}}(k)$, encompassing both stationary and transient signal frames, is shown in Figure 2.5. In the example, four 20 ms frames are classified as transient.

As a possible extension, the length and the shape of the interpolation window $w_\mathrm{T}(k)$ in (2.6) can also be adapted to the frame type (transient or stationary). This method, aiming to reduce spectral leakage further, has been implemented in the codec proposal of [Geiser et al. 2009], see also Section 3.3. However, for simplicity, the basic interpolation method of (2.6) shall be used exclusively in the present chapter.

2.2.3 Other Applications for Temporal Envelope Control

Apart from bandwidth extension, the proposed mechanism for temporal envelope control of audio signals can also be used for other applications, namely frame erasure concealment (FEC) and audio coding. This is summarized below.

Frame Erasure Concealment

Frame erasure concealment (FEC) is an essential component of audio codecs that are being used in packet switched network environments. Therefore, FEC modules have recently been added to several standardized codecs such as ITU-T G.722 [ITU-T 1984]. New codecs, e.g., ITU-T G.729.1 [ITU-T 2006, Ragot et al. 2007] or ITU-T G.718 [ITU-T 2008*a*, Vaillancourt et al. 2008], are directly designed under such constraints. Typically, the encoder adds a certain amount of dedicated FEC information to the bitstream. For example, the transmitted FEC side information might comprise a coarse but relevant description of *past* signal frames, in particular their *energy envelope* but also, e.g., rough phase information (for strongly periodic signals). Therefore, if a frame has been lost during transmission, an approximate version can still be reproduced based on the decoder's memory and on the available FEC information.

A concealment which is based on information about past signal frames requires at least a *one-frame delay* at the decoder side in order to handle (single) frame losses. As a matter of fact, audio codecs based on frequency transforms using half-overlapped windowing (lapped transforms, e.g., the popular MDCT, see also Section 2.4.3) inherently incur a one-frame delay within the decoder because of the required overlap-add operation. This property can be elegantly exploited to transmit FEC information for single frame losses *without any additional delay-penalty* provided that the computation of this information does not use additional look-ahead samples. An example realization which extracts the FEC side information directly from the time domain signal is used in the ITU-T G.729.1 codec [ITU-T 2006, Ragot et al. 2007].

Also the temporal envelope control scheme as described in Section 2.2.1 facilitates the reuse of information for FEC in the extension band. Therefore, a special variable bitstream arrangement can be used for a given frame λ. Concretely, the encoded subframe gains with *odd* indices λ_{SF} from the *current* frame (i.e., $g(\lambda, \lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{TE}} - 1\}$) and the encoded subframe gains with *even* indices λ_{SF} from the *previous* frame (i.e., $g(\lambda - 1, \lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{0, 2, \ldots, N_{\mathrm{TE}} - 2\}$) are transmitted in the *same* bitstream frame (packet). Without frame erasures, both gain subsets are available at the decoder due to the one-frame delay. However, if a *frame erasure* is signaled to the decoder, the (single) received subset (even or odd indices λ_{SF}) can be interpolated to form a temporal gain function of lower resolution. If, moreover, an *adaptive* temporal envelope method with transient/stationary classification according to Section 2.2.2 is applied, signalization flags and gains in *stationary* signal segments are transmitted redundantly for con-

cise results. A detailed proposal for FEC of a 8 – 14 kHz extension band signal which is based on an adaptive temporal envelope representation is described in [Geiser & Vary 2009]. This algorithm has been used in the codec proposal of [Geiser et al. 2009], see also Section 3.3.

Temporal Envelope Control in Multiple Spectral Bands

The previously discussed methods for temporal gain control are designed to operate on subband signals. However, enforcing a temporal gain contour for the *full* frequency band (e.g., 0 – 8 kHz) might result an unnatural and "snatchy" sound character. The plain gain control method from above is therefore only applicable with subband algorithms that already provide an inherent subband decomposition as, e.g., the present BWE system (Figure 2.1). The algorithm can, however, be generalized to a multi-band scenario [Geiser, Roggendorf & Vary 2010] where a *uniform or non-uniform frequency selectivity* is achieved by applying the concept of a *filterbank equalizer* (FBE) [Vary 2006, Löllmann & Vary 2007]. This multi-band technique is suitable for an application in audio codecs where, in particular, pre-echo control or an adaptive spectro-temporal pre- and deemphasis can be realized.[1]

2.3 Autoregressive Representation of the Spectral Envelope

The most common *spectral envelope* representation is based on the *autoregressive* (AR) signal model. This approach is well-known from speech codecs that employ *linear predictive coding* (LPC) techniques. The AR model is also used in numerous bandwidth extension algorithms that have been proposed in the literature, e.g., [Carl & Heute 1994, Jax & Vary 2003].

2.3.1 Analysis

The LPC parameters (or AR coefficients) for the λ-th signal frame are obtained by fitting an all-pole model of order N_{SE} to a windowed segment $s^{\mathrm{w}}_{\mathrm{eb}}(k) = s_{\mathrm{eb}}(k + \lambda L) \cdot w_{\mathrm{LP}}(k)$ of the input signal $s_{\mathrm{eb}}(k)$.[2] Thereby, $w_{\mathrm{LP}}(k)$ for $k \in \{0, \ldots, L_{\mathrm{w}} - 1\}$ is the analysis window. Hence, the AR coefficient vector $\mathbf{a}(\lambda) = (a_1(\lambda), a_2(\lambda), \ldots, a_{N_{\mathrm{SE}}}(\lambda))^{\mathrm{T}}$ is defined as

$$\mathbf{a}(\lambda) = \arg \min_{(a_1, \ldots, a_{N_{\mathrm{SE}}})^{\mathrm{T}}} \mathrm{E}\Big\{ \Big(s^{\mathrm{w}}_{\mathrm{eb}}(k) - \sum_{j=1}^{N_{\mathrm{SE}}} a_j \cdot s^{\mathrm{w}}_{\mathrm{eb}}(k-j) \Big)^2 \Big\}. \tag{2.11}$$

Following the well-known autocorrelation method, e.g., [Vary & Martin 2006], the coefficients $a_j(\lambda)$ are in practice found by solving the *Yule-Walker* (or *normal*)

[1] The FBE concept as such is also applicable to the bandwidth extension problem. This is discussed in Section 2.4.4.

[2] In the *serial* analysis approach of Figure 2.3(a), $s^{\mathrm{t}}_{\mathrm{eb}}(k)$ is used as the input signal instead.

equations

$$\begin{pmatrix} \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(1) \\ \vdots \\ \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(N_{\mathrm{SE}}) \end{pmatrix} = \begin{pmatrix} \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(0) & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(-1) & \dots & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(1-N_{\mathrm{SE}}) \\ \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(1) & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(0) & \dots & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(2-N_{\mathrm{SE}}) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(N_{\mathrm{SE}}-1) & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(N_{\mathrm{SE}}-2) & \dots & \hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(0) \end{pmatrix} \begin{pmatrix} a_1(\lambda) \\ \vdots \\ a_{N_{\mathrm{SE}}}(\lambda) \end{pmatrix} \tag{2.12}$$

whereby $\hat{\varphi}_{s^{\mathrm{w}}_{\mathrm{eb}} s^{\mathrm{w}}_{\mathrm{eb}}}(\kappa)$ is a short-term estimate (for frame λ) of the autocorrelation function of $s^{\mathrm{w}}_{\mathrm{eb}}(k)$. This equation system is efficiently solved using *Levinson-Durbin* recursions, cf. [Vary & Martin 2006]. The obtained AR coefficients $a_j(\lambda)$ can then be used to compute the *linear prediction residual*:

$$u^{\mathrm{LP}}_{\mathrm{eb}}(k) = s_{\mathrm{eb}}(k) - \sum_{j=1}^{N_{\mathrm{SE}}} a_j(\lambda) \cdot s_{\mathrm{eb}}(k-j). \tag{2.13}$$

2.3.2 Synthesis with Artificial Excitation Signals

Based on the received (or estimated) AR coefficients $\hat{\mathbf{a}}(\lambda) = (\hat{a}_1(\lambda), \dots, \hat{a}_{N_{\mathrm{SE}}}(\lambda))^{\mathrm{T}}$, signal synthesis can be performed by the receiver:

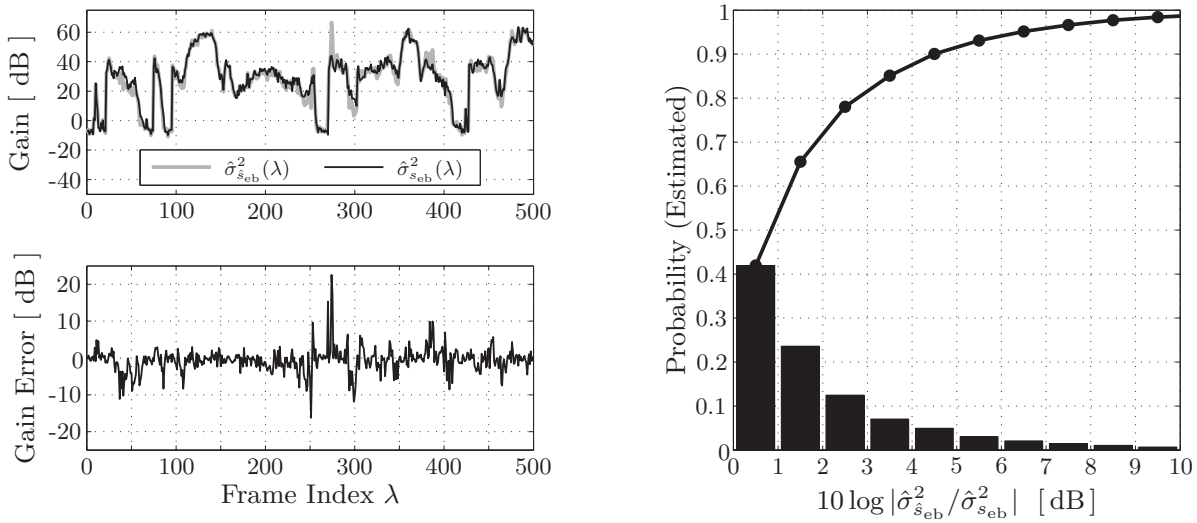
$$\hat{s}_{\mathrm{eb}}(k) = \hat{u}_{\mathrm{eb}}(k) + \sum_{j=1}^{N_{\mathrm{SE}}} \hat{a}_j(\lambda) \cdot \hat{s}_{\mathrm{eb}}(k-j) \tag{2.14}$$

whereby the signal $\hat{u}_{\mathrm{eb}}(k)$ is called the "excitation" signal of the synthesis filter (see Section 2.5). If $\hat{\mathbf{a}}(\lambda) = \mathbf{a}(\lambda)$ and $\hat{u}_{\mathrm{eb}}(k) = u^{\mathrm{LP}}_{\mathrm{eb}}(k)$, the original input signal $s_{\mathrm{eb}}(k)$ is exactly reconstructed[3] by (2.14). Lossy speech codecs transmit a *quantized* LPC residual which guarantees an approximate reconstruction of the input signal. However, in bandwidth extension algorithms, an *artificial* excitation $\hat{u}_{\mathrm{eb}}(k)$ is used instead of a quantized signal (see Section 2.5). Such artificial signals are in general not directly related to the original LPC residual in (2.13). Sometimes only the (approximate) gain is enforced for the excitation signal, i.e.,

$$\hat{\sigma}^2_{\hat{u}_{\mathrm{eb}}}(\lambda) \stackrel{!}{=} \hat{\sigma}^2_{u^{\mathrm{LP}}_{\mathrm{eb}}}(\lambda) \doteq \frac{1}{L-1} \sum_{j=0}^{L-1} \left[u^{\mathrm{LP}}_{\mathrm{eb}}(j + \lambda \cdot L) \right]^2. \tag{2.15}$$

However, even with matched gains according to (2.15), it is not ensured that $\hat{\sigma}^2_{\hat{s}_{\mathrm{eb}}}(\lambda) = \hat{\sigma}^2_{s_{\mathrm{eb}}}(\lambda)$, i.e., the (short term) power of the synthesized signal $\hat{s}_{\mathrm{eb}}(k)$ may deviate from the (short term) power of $s_{\mathrm{eb}}(k)$. This effect is illustrated in Figure 2.6. In the experiment, the LP residual of the 0 – 4 kHz band of a speech signal is used as the excitation signal $\hat{u}_{\mathrm{eb}}(k)$ to synthesize the 4 – 8 kHz extension band $\hat{s}_{\mathrm{eb}}(k)$. The extension band gain $\hat{\sigma}^2_{\hat{s}_{\mathrm{eb}}}(\lambda)$ after AR synthesis has been measured for

[3] An adequate implementation of the adaptive filters is assumed.



(a) Extension band gains and gain error for an example signal segment.

(b) Histogram (■) and cumulative density function (CDF) (–●–), collected over 1 min of speech.

Figure 2.6: Gain mismatch in AR synthesis with artificial excitation.

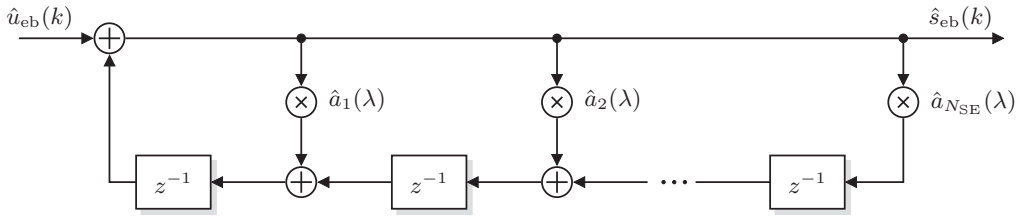
each frame. In the figure, $\hat{\sigma}^2_{\hat{s}_{\mathrm{eb}}}(\lambda)$ is compared with the reference (original) gain of $s_{\mathrm{eb}}(k)$. It can be observed that, in certain critical signal segments, significant deviations from the reference gain contour occur in the synthesized signal.

As a solution, excitation signals of *unit* gain ($\hat{\sigma}^2_{\hat{u}_{\mathrm{eb}}}(\lambda) = 1$) can be used in (2.14) in combination with an *external* gain control for the *synthesized* signal $\hat{s}_{\mathrm{eb}}(k)$. Here, this is achieved with the temporal envelope shaping method proposed in Section 2.2. Moreover, with the gain shaping in the signal domain instead of the residual domain, the explicit computation of the extension band LP residual according to (2.13) is no longer required for the determination of the reference parameters $\mathbf{p}(\lambda)$ and can be omitted.

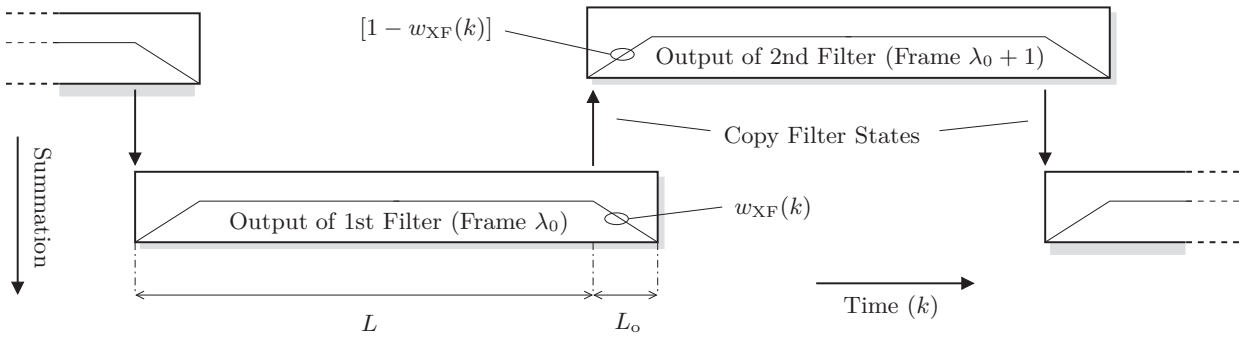
2.3.3 Synthesis Filter Implementation

An important aspect for the implementation of AR synthesis (2.14) is the employed filter structure and the filter adaptation mechanism. The most common choice is an implementation in direct or transposed form while the set of filter coefficients $\mathbf{a}(\lambda)$ is instantaneously switched at the frame boundaries ($k = \lambda L$). The *transposed* form of the AR synthesis filter has the advantage that the new coefficient set is not immediately effective at the filter output since the weighted samples are buffered before the output, see Figure 2.7(a). This causes a certain smoothing effect at the frame boundaries which is particularly important if an artificial excitation signal $\hat{u}_{\mathrm{eb}}(k)$ is applied to the synthesis filter.

The frame transitions can be smoothed further if a "crossfading" method is applied. The idea here is to define an overlap period $L_{\mathrm{o}} \leq L$ in which, effectively,



(a) AR synthesis filter in transposed direct form.



(b) Filter crossfading method.

Figure 2.7: Implementation of the AR synthesis filter.

two parallel synthesis filters are applied. Both filter outputs are then combined with a suitable crossfading function. Consider, as illustrated in Figure 2.7(b), the transition from frame λ_0 to frame $(\lambda_0 + 1)$. The output of the first filter is then

$$\hat{s}_1(k) = \hat{u}_{\mathrm{eb}}(k) + \sum_{j=1}^{N_{\mathrm{SE}}} \hat{a}_j(\lambda_0) \cdot \hat{s}_1(k-j) \tag{2.16}$$

for $k \in \{\lambda_0 L, \ldots, (\lambda_0 + 1)L + L_{\mathrm{o}} - 1\}$. The output for frame $(\lambda_0 + 1)$ is computed for $k \in \{(\lambda_0 + 1)L, \ldots, (\lambda_0 + 2)L + L_{\mathrm{o}} - 1\}$ as follows:

$$\hat{s}_2(k) = \hat{u}_{\mathrm{eb}}(k) + \sum_{j=1}^{N_{\mathrm{SE}}} \hat{a}_j(\lambda_0 + 1) \cdot \hat{s}_2(k-j). \tag{2.17}$$

During the crossfading period, i.e., for $k \in \{(\lambda_0+1)L, \ldots, (\lambda_0+1)L + L_{\mathrm{o}} - 1\}$, the crossfaded signal is obtained as

$$\hat{s}_{\mathrm{eb}}(k) = w_{\mathrm{XF}}(k - (\lambda_0 + 1)L) \cdot \hat{s}_1(k) + [1 - w_{\mathrm{XF}}(k - (\lambda_0 + 1)L)] \cdot \hat{s}_2(k) \tag{2.18}$$

with the crossfading function $w_{\mathrm{XF}}(k)$. This crossfading function can be chosen as the falling slope of any common window function, e.g., the triangular window or the Hann window. Furthermore, to ensure the continuity of the filtering process, the *filter states* have to be taken into account. For the described crossfading

mechanism, the states of the second filter (which starts to operate at sample index $(\lambda_0 + 1) \cdot L$) are initialized by duplicating the states from the first filter before the output sample $\hat{s}_1((\lambda_0 + 1) \cdot L)$ is produced. Regarding the *filter structure*, the same comments apply as for the conventional adaptive filtering method, i.e., the *transposed filter implementation* from Figure 2.7(a) is advisable.

The crossfading method implies a processing delay of $L_\mathrm{o}/2$ samples because of the overlap-add operation (2.18). Moreover, it has to be noted that, instead of crossfading the output signals $\hat{s}_1(k)$ and $\hat{s}_2(k)$ of two separate filters, also the filter *coefficients* $\hat{a}_j(\lambda)$ could be crossfaded. However, this operation is only equivalent to the described mechanism if finite impulse response (FIR) filters are used, whereas the AR synthesis of (2.14) is a recursive filter and therefore has an infinite impulse response (IIR).

2.4 Spectral Envelope Modeling in the Frequency Domain

As an alternative to AR modeling of the spectral envelope, it can be desirable to derive spectral envelope parameters from a frequency domain representation of the audio signal. This is particularly relevant when the parametric signal synthesis shall be used within the framework of an *existing* transform based audio codec.

2.4.1 Spectral Transforms

The frequency domain representation for frame λ of the real-valued signal $s_\mathrm{eb}(k)$ is denoted by $S_\mathrm{eb}(\lambda, \mu)$ with the index μ for the frequency bins. The most common frequency transform is the *Discrete Fourier Transform* (DFT)

$$S_\mathrm{eb}^\mathrm{DFT}(\lambda, \mu) = \sum_{k=0}^{L_\mathrm{w}-1} w_\mathrm{F}(k)\, s_\mathrm{eb}(k + \lambda L) \cdot \mathrm{e}^{-\jmath \frac{2\pi\mu k}{L_\mathrm{w}}} \tag{2.19}$$

with $\mu \in \left\{0, \ldots, \frac{L_\mathrm{w}}{2}\right\}$. Note that the symmetric extension for the transform coefficients of real input signals, i.e., $S_\mathrm{eb}^\mathrm{DFT}(\lambda, \frac{L_\mathrm{w}}{2} + \mu) = S_\mathrm{eb}^\mathrm{DFT}(\lambda, \frac{L_\mathrm{w}}{2} - \mu)^*$ with $\mu \in \left\{1, \ldots, \frac{L_\mathrm{w}}{2} - 1\right\}$, is omitted here. The employed analysis window (or prototype filter) $w_\mathrm{F}(k)$ has a length of L_w samples.

Instead of the DFT, modern transform audio codecs typically use the half-overlapped (i.e., $L_\mathrm{w} \overset{!}{=} 2\,L$), critically sampled, and real-valued *Modified Discrete Cosine Transform* (MDCT) [Princen & Bradley 1986, Malvar 1992]

$$S_\mathrm{eb}^\mathrm{MDCT}(\lambda, \mu) = \sum_{k=0}^{L_\mathrm{w}-1} w_\mathrm{F}(k)\, s_\mathrm{eb}(k + \lambda L) \cdot \cos\left[\frac{\pi}{L}\left(k + \frac{L+1}{2}\right)\left(\mu + \frac{1}{2}\right)\right] \tag{2.20}$$

with $\mu \in \left\{0, \ldots, \frac{L_\mathrm{w}}{2} - 1\right\}$, whereby, in this case, the window $w_\mathrm{F}(k)$ has to fulfill the Princen-Bradley conditions [Princen & Bradley 1986] to achieve a perfect reconstruction after inverse transform and overlap-add.

2.4.2 Subband Gains

Based on a given frequency transform representation $S_{\mathrm{eb}}(\lambda, \mu)$ for frame λ, a fixed number N_{SE} of *subband gains* with index $m \in \{0, \ldots, N_{\mathrm{SE}} - 1\}$ can be defined (in analogy to the subframe gains from Section 2.2):

$$\gamma(\lambda, m) = \max \left\{ \gamma_0, \sqrt{\frac{1}{M_{\mathrm{SB}}} \sum_{\mu=0}^{M_{\mathrm{SB}}-1} W(\mu) \cdot |S_{\mathrm{eb}}(\lambda, \mu + mM_{\mathrm{s}})|^2} \right\}. \tag{2.21}$$

Thereby, M_{SB} is the subbandwidth (in bins) and $\gamma_0 > 0$ is, again, a fixed minimum gain value. The frequency domain window $W(\mu)$ spans a bandwidth of M_{SB} frequency bins. The subband spacing is $M_{\mathrm{s}} \leq M_{\mathrm{SB}}$ bins, i.e., there is an optional spectral overlap of $M_{\mathrm{SB}} - M_{\mathrm{s}}$ frequency bins. The number N_{SE} of spectral gains is associated with the subband spacing M_{s}, with the transform length L_{w} and with the desired cutoff frequency f_{c} via

$$N_{\mathrm{SE}} = \left\lceil \frac{L_{\mathrm{w}} f_{\mathrm{c}}}{M_{\mathrm{s}} f_{\mathrm{s}}} \right\rceil. \tag{2.22}$$

It should be noted that, as an alternative to (2.21), spectral subband gains can be obtained immediately from the time domain signal with the help of a polyphase DFT analysis filterbank with N_{SE} channels [Geiser, Roggendorf & Vary 2010]. However, this case is not considered here since, in the context of a transform codec, the frequency transform and the transform coefficients $S_{\mathrm{eb}}(\lambda, \mu)$ are assumed to be given.[4]

The spectral gains of (2.21) are, in the general case, computed with *overlapping* windows $W(\mu)$. Yet, in typical applications in transform audio codecs, *rectangular* windowing without any overlap is desirable, i.e., $M_{\mathrm{SB}} = M_{\mathrm{s}}$ and $W(\mu) \equiv 1$. The quantized gains $\hat{\gamma}(\lambda, m)$ can then be reused as scalefactors for spherical vector quantization of transform coefficients as employed in many codecs, see Section 3.3.

2.4.3 Signal Synthesis in the Frequency Domain

A frequency domain representation of the spectral envelope is particularly useful if the bandwidth extension algorithm shall be tightly integrated with a transform audio codec. Then, the frequency domain is also a natural choice for spectral envelope *synthesis* in the decoder.

Spectral Gain Function

The synthesis begins with a (spectrally flat) "excitation signal" $\hat{U}_{\mathrm{eb}}(\lambda, \mu)$, see Section 2.5, based on which spectral correction gains can be established:

$$\hat{\gamma}_{\mathrm{rel}}(\lambda, m) = \frac{\hat{\gamma}(\lambda, m)}{\hat{\gamma}'(\lambda, m)}. \tag{2.23}$$

[4] The corresponding time domain *synthesis* algorithm is based on the concept of a filterbank equalizer (FBE) [Vary 2006, Löllmann & Vary 2007]. This is discussed in Section 2.4.4.

In (2.23), $\hat{\gamma}(\lambda, m)$ are the received or estimated spectral parameters and $\hat{\gamma}'(\lambda, m)$ are the *measured* spectral gains (according to (2.21)) of $\hat{U}_{\mathrm{eb}}(\lambda, \mu)$. Then, a frequency domain equivalent for the temporal gain function (TGF) of Section 2.2, i.e., a *spectral gain function* (SGF), can be defined

$$\begin{aligned}
&\hat{\gamma}_{\mathrm{SGF}}(\lambda, \mu + mM_{\mathrm{s}}) \\
&\quad = \begin{cases}
W_{\mathrm{S}}(\mu + \frac{M_{\mathrm{s}}}{2}) \cdot \hat{\gamma}_{\mathrm{rel}}(\lambda, m) + \left[1 - W_{\mathrm{S}}(\mu + \frac{M_{\mathrm{s}}}{2})\right] \cdot \hat{\gamma}_{\mathrm{rel}}(\lambda, m-1) \\
\qquad\qquad \text{if} \quad \mu \in \{0, \ldots, M_{\mathrm{s}}/2 - 1\} \\
[1 - W_{\mathrm{S}}(\mu)] \cdot \hat{\gamma}_{\mathrm{rel}}(\lambda, m) \;\; + W_{\mathrm{S}}(\mu) \cdot \hat{\gamma}_{\mathrm{rel}}(\lambda, m+1) \\
\qquad\qquad \text{if} \quad \mu \in \{M_{\mathrm{s}}/2, \ldots, M_{\mathrm{s}} - 1\}
\end{cases}
\end{aligned} \tag{2.24}$$

with the interpolation window $W_{\mathrm{S}}(\mu)$ which, in most cases, should be matched with the window function $W(\mu)$ of the analysis side. To restore the spectral signal characteristics, the SGF is applied to the excitation signal:

$$\hat{S}_{\mathrm{eb}}(\lambda, \mu) = \hat{U}_{\mathrm{eb}}(\lambda, \mu) \cdot \hat{\gamma}_{\mathrm{SGF}}(\lambda, \mu). \tag{2.25}$$

According to Figure 2.3(c), an eventual *temporal* impact of the spectral multiplication is subsequently corrected by the application of the temporal gain function $\hat{g}_{\mathrm{TGF}}(k)$ as described in Section 2.2.1.

Inverse Transform

The spectrally shaped signal $\hat{S}_{\mathrm{eb}}(\lambda, \mu)$ for each frame λ is finally transformed to the time domain. For the case of the *Inverse Discrete Fourier Transform* (IDFT), the frequency domain symmetry conditions for real-valued time domain signals have to be considered, i.e., the inverse transform equation becomes

$$\tilde{s}_{\mathrm{eb}}^{\mathrm{DFT}}(k + \lambda L) = \frac{1}{L} \sum_{\mu=0}^{\frac{L_{\mathrm{w}}}{2}-1} \left[\hat{S}_{\mathrm{eb}}^{\mathrm{DFT}}(\lambda, \mu) + (-1)^k \cdot \hat{S}_{\mathrm{eb}}^{\mathrm{DFT}}\left(\lambda, \mu + \frac{L_{\mathrm{w}}}{2}\right)\right] \cdot \mathrm{e}^{\jmath \frac{2\pi\mu k}{L_{\mathrm{w}}}} \tag{2.26}$$

with $k \in \{0, \ldots, L_{\mathrm{w}} - 1\}$. The counterpart to the MDCT of (2.20) is the *Inverse Modified Discrete Cosine Transform* (IMDCT):

$$\tilde{s}_{\mathrm{eb}}^{\mathrm{MDCT}}(k, \lambda) = \frac{1}{L} \sum_{\mu=0}^{L-1} \hat{S}_{\mathrm{eb}}^{\mathrm{MDCT}}(\lambda, \mu) \cos\left[\frac{\pi}{L}\left(k + \frac{L+1}{2}\right)\left(\mu + \frac{1}{2}\right)\right] \tag{2.27}$$

with $k \in \{0, \ldots, L_{\mathrm{w}} - 1\}$. In both cases (DFT and MDCT), an overlap-add operation has to be carried out and the final time domain signal is determined as

$$\hat{s}_{\mathrm{eb}}(k + \lambda L) = w_{\mathrm{F}}(k) \cdot \tilde{s}_{\mathrm{eb}}(k, \lambda) + w_{\mathrm{F}}(k + L) \cdot \tilde{s}_{\mathrm{eb}}(k, \lambda - 1) \tag{2.28}$$

for $k \in \{0, \ldots, L\}$ and $\lambda \in \mathbb{Z}$. In the IMDCT case, this operation is particularly important since it cancels out the time domain alias, cf. [Princen & Bradley 1986].

Comments on Parametric Signal Analysis/Synthesis in the MDCT Domain

The MDCT is very well suited for the efficient quantization of spectral coefficients in audio codecs because it simultaneously offers *critical downsampling*, *overlapped framing*, and *perfect reconstruction*. However, the latter is only achieved by the final overlap-add step in the time domain (2.28), i.e., through time domain alias cancellation (TDAC). Hence, the $L_\mathrm{w}/2$ real-valued transform coefficients for a *single* frame with index λ do *not* convey the full information on the L_w time domain samples of the corresponding input window. As a particular consequence, Parseval's theorem does not hold for the MDCT [Wang et al. 2000] and energy calculations (such as (2.21)) that are based on this representation are not exact. Therefore, a dedicated *complex valued* filterbank that facilitates a more reliable parameter extraction is often used instead, e.g., in the "spectral band replication" technique of [Dietz et al. 2002]. However, despite certain disadvantages for concise parameter estimation, the MDCT still proved to be useful for parametric coding because a seamless integration with transform audio codecs is possible, e.g., [Geiser et al. 2009, Tammi et al. 2009, Tsujino & Kikuiri 2009, Laaksonen et al. 2010]. For instance, the energy parameters from MDCT subbands can be reused as scale factors for MDCT domain spherical vector quantization, e.g., [Geiser et al. 2009].

2.4.4 Signal Synthesis in the Time Domain

An alternative to the frequency domain synthesis approach as outlined in the previous section is a filter-based *time domain synthesis* approach using the received (or estimated) spectral gains $\hat{\gamma}(\lambda, \mu)$. A suitable tool to map the gains $\hat{\gamma}(\lambda, \mu)$ to the coefficients $h_\mathrm{FBE}(\lambda, k)$ of a time domain filter is a *filterbank equalizer* (FBE) [Vary 2006]. The typical use case for FBEs is speech enhancement, in particular noise reduction [Löllmann & Vary 2007] and near end listening enhancement [Sauert et al. 2008]. An application to pre-echo control in audio coding is discussed in [Geiser, Roggendorf & Vary 2010]. Also, a few proposals for bandwidth extension algorithms based on FBE and related techniques have been made in the literature, e.g., [Geiser et al. 2007a, Kim et al. 2008, Pulakka et al. 2010].

The FBE frequency channels with index $m \in \{0, ..., N_\mathrm{SE} - 1\}$ are defined by their bandpass impulse response $h_\mathrm{FBE}^{(m)}(k)$. These subband filters are modulated versions of the linear phase FIR prototype low-pass filter $h_0(k)$ of length L_FBE which, for perfect reconstruction, is required to fulfill the so called "M-th band condition," i.e., $h_0(nM + L_\mathrm{FBE}/2) = 0$ for $n \in \mathbb{Z} \setminus \{0\}$ and $M = \frac{L_\mathrm{w}}{2M_\mathrm{s}}$, see e.g., [Mintzer 1982]. Based on the prototype design, individual bandpass filters are then derived as modulated versions of the prototype impulse response. In [Vary 2006] and [Löllmann & Vary 2007], the modulation is effectively obtained with a generalized DFT (GDFT) of the time varying spectral gains (weights) and by a subsequent multiplication with the impulse response of the prototype low-pass filter. Here, in contrast, the subband filters are obtained by real-valued cosine

modulation resulting in subband filters with linear phase:

$$h_{\mathrm{FBE}}^{(m)}(k) = h_0(k) \cdot \cos\left((m + m_0) \cdot \frac{2\pi M_{\mathrm{s}} k}{L_{\mathrm{w}}}\right). \tag{2.29}$$

A frequency shift of $m_0 = 1/2$ is applied to match the subband definition of (2.21). The overall FBE impulse response $h_{\mathrm{FBE}}(\lambda, k)$ for each frame λ is finally obtained by the weighted sum

$$h_{\mathrm{FBE}}(\lambda, k) = \sum_{m=0}^{N_{\mathrm{SE}}-1} \hat{\gamma}_{\mathrm{rel}}(\lambda, m) \cdot h_{\mathrm{FBE}}^{(m)}(k) \tag{2.30}$$

where $\hat{\gamma}_{\mathrm{rel}}(\lambda, m)$ can, for instance, be determined according to (2.23). The filter equation for the FBE is then

$$\hat{s}_{\mathrm{eb}}(k + \lambda L) = \sum_{j=0}^{L_{\mathrm{FBE}}-1} \hat{u}_{\mathrm{eb}}(k - j) \cdot h_{\mathrm{FBE}}(\lambda, k) \tag{2.31}$$

for $k \in \{0, \ldots, L-1\}$. To accommodate heavily time-varying gains (or filter coefficients), either the filter crossfading method of Section 2.3.3 and Figure 2.7(b) can be reused or, since (2.31) specifies an FIR filter, a direct interpolation of the correction gains $\hat{\gamma}_{\mathrm{rel}}(\lambda, m)$ can be applied.

The FBE with N_{SE} channels can be designed such that the individual filterbank channels match the corresponding analysis frequency subbands in (2.21) as closely as possible. In particular, the filter prototype and the modulation frequencies have to be chosen accordingly. In this case, also overlapped frequency domain analysis windows $W(\mu)$ with $M_{\mathrm{s}} \leq M_{\mathrm{SB}}$ can be interesting to align both the FBE prototype bandwidth and the FBE intra-channel overlap.

Compared to the frequency domain synthesis approach of Section 2.4.3, the FBE concept offers several advantages. For example, with the FBE synthesis, a lower algorithmic delay and a reduced computational complexity can be achieved, both because of the fact that the inverse transform can be omitted. Moreover, there is the possibility to interchange the time and frequency envelope shaping blocks in Figure 2.3(c). However, the filtering operations of (2.31) may influence the *temporal* signal characteristics in an undesired way (similar to the *spectral* leakage effects as discussed in Section 2.2.1). Potentially, the temporal energy distribution is "smeared" over an interval which corresponds to the length of the frequency envelope shaping filter (i.e., L_{FBE} taps). However, with relatively wide and overlapping frequency responses of the filterbank channels, it is guaranteed that essential temporal characteristics are maintained, see [Geiser et al. 2007a]. An FBE based synthesis of the spectral envelope is used in the TDBWE algorithm of ITU-T G.729.1. The related filterbank design is described in more detail in the following chapter (Section 3.2.3). The corresponding amplitude transfer functions of the individual filters are then shown in Figure 3.8.

2.5 Parametric Regeneration of Spectral Details

In the simplest case, the term "spectral details" refers to the linear prediction residual $u_{\mathrm{eb}}^{\mathrm{LP}}(k)$ as defined by (2.13). However, in the context of bandwidth extension, "spectral details" are interpreted more generally, i.e., as the (artificially generated) signal $\hat{u}_{\mathrm{eb}}(k)$ which is used to drive the subsequent envelope shaping components of the algorithm, see Figure 2.3(c). In fact, a large number of methods exists to regenerate the spectral details $\hat{u}_{\mathrm{eb}}(k)$. These approaches range from simple pseudo-random noise generators over signal processing techniques to derive $\hat{u}_{\mathrm{eb}}(k)$ from the received baseband signal $\hat{s}_{\mathrm{bb}}(k)$ (e.g., spectral replication or spectral folding) to sophisticated parameter-driven methods which are based on the modeling of individual harmonic components or pitch cycles. Hybrid algorithms are also possible. The methods can be categorized as "blind" (i.e., no side information transmitted) or "non-blind" (i.e., side information is available).

This section briefly summarizes several common methods to regenerate the spectral details of the extension band at the receiver side. Two *concrete* algorithms, which have been designed for particular applications, are detailed in Chapter 3.

2.5.1 Spectral Replication

Spectral replication is a simple yet effective way to generate the excitation signal either in the time domain ($\hat{u}_{\mathrm{eb}}(k)$) or in the frequency domain ($\hat{U}_{\mathrm{eb}}(\lambda,\mu)$) based on the respective baseband signal $\hat{s}_{\mathrm{bb}}(k)$ or $\hat{S}_{\mathrm{bb}}(\lambda,\mu)$. Spectral replication methods reuse a "spectrally flattened" baseband signal as the extension band excitation. The method can be realized with or without side information.

Early proposals for spectral replication date back to [Un & Magill 1975] and [Makhoul & Berouti 1979]. Also the GSM FullRate codec [ETSI 1990, Vary et al. 1988] is based on spectral replica of a quantized baseband signal that is 1.66 kHz wide. Interestingly, the replication method is not only useful for bandwidth extension of speech signals, but it is also applicable for extending the bandwidth of audio material. The approach is meanwhile successfully applied in many codec standards, for instance in the "Spectral Band Replication" (SBR) tool in MPEG Audio Coding [Dietz et al. 2002].

Frequency Domain Implementation

Spectral replication in the frequency domain is relatively straight forward. First, the baseband spectrum $S_{\mathrm{bb}}(\lambda,\mu)$ in frame λ is processed to obtain a spectrally flat signal. This can, for instance, be achieved by spectral normalization with the spectral gain function of $S_{\mathrm{bb}}(\lambda,\mu)$ according to (2.24)

$$S_{\mathrm{bb}}^{\mathrm{norm}}(\lambda,\mu) = S_{\mathrm{bb}}(\lambda,\mu)\cdot\gamma_{\mathrm{SGF}}^{-1}(\lambda,\mu). \tag{2.32}$$

Then, a suitable part of $S_{\mathrm{bb}}^{\mathrm{norm}}(\lambda,\mu)$ is reused as excitation signal. Speech signals are often less tonal in the extension band. Therefore, a certain amount of pseudo random noise $N(\lambda,\mu)$ can be added optionally.

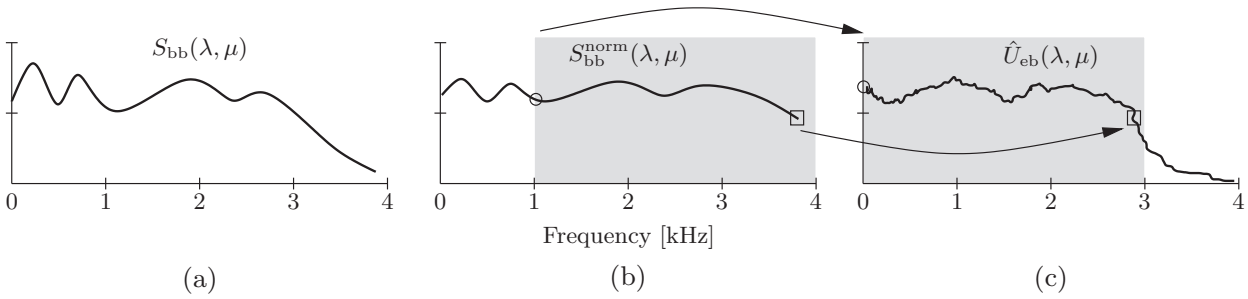


Figure 2.8: Spectral replication in the frequency domain.

As an example, an excitation signal for a bandwidth extension from 4 kHz (NB) to 7 kHz (WB) shall be considered, see Figure 2.8. In this case, both the baseband and the extension band signals are sampled at 8 kHz. If the MDCT (2.20) is assumed as frequency transform, there are $L_{\mathrm{w}}/2$ real-valued frequency bins for the baseband as well as the extension band. Then, in the example, the desired excitation signal is synthesized as

$$\hat{U}_{\mathrm{eb}}(\lambda,\mu) = \begin{cases} S_{\mathrm{bb}}^{\mathrm{norm}}(\lambda,\mu+\frac{1}{8}\cdot L_{\mathrm{w}}) + N(\lambda,\mu) & \text{for } \mu \in \left\{0,\ldots,\frac{3L_{\mathrm{w}}}{8}-1\right\} \\ 0 & \text{for } \mu \in \left\{\frac{3L_{\mathrm{w}}}{8},\ldots,\frac{L_{\mathrm{w}}}{2}-1\right\}. \end{cases} \tag{2.33}$$

Note that the 1 – 4 kHz range of the baseband spectrum (offset of $L_{\mathrm{w}}/8$ bins) is used as excitation for the 4 – 7 kHz range (frequency bins 0 to $3L_{\mathrm{w}}/8-1$ in the extension band MDCT domain). This choice is particularly beneficial for speech signals since the 0 – 1 kHz band often contains very strong harmonics which are atypical for the extension band.

The blind replication method of (2.33) can also be supported by a limited amount of side information. In general, this side information has to be extracted from the *original* extension band spectrum $S_{\mathrm{eb}}(\lambda,\mu)$. The following options can be considered:

- Tonality adjustment — The tonality of the replicated signal can be adjusted according to the measured tonality of the original extension band signal. Concretely, "peak sharpening" for too noisy signals or adaptive "noise mixing" for too tonal signals can be employed, see e.g., [Geiser et al. 2009].

- Pitch cutoff frequency — A frequency index μ_{c} is determined. Below this cutoff frequency, the usual replication method of (2.33) is applied. For higher frequencies, i.e., for $\mu \geq \mu_{\mathrm{c}}$, more emphasis is put on the noise contribution $N(\lambda,\mu)$. This way, the natural voicing characteristics of human speech can be reproduced more accurately.

- Spectral "patching" — The spectral replication method is in principle not restricted to a single continuous frequency range. More flexibility is added by allowing spectral "patches" that are copied from the baseband spectrum

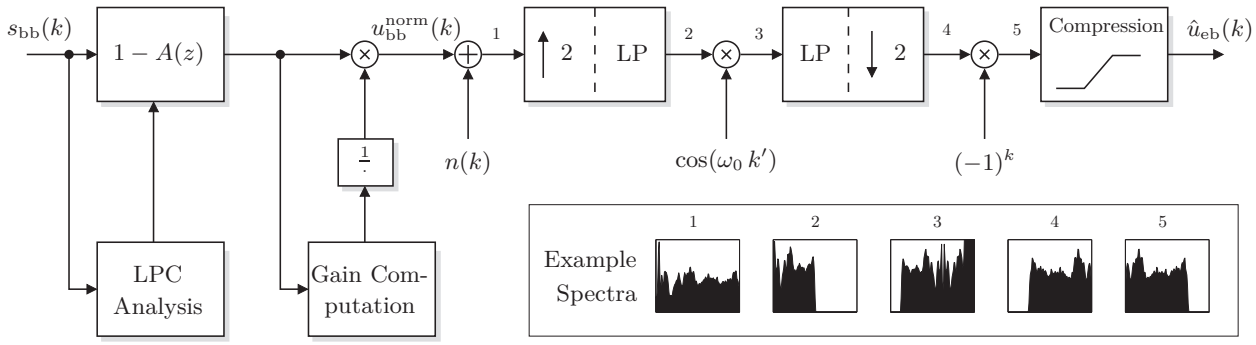


Figure 2.9: System for spectral replication in the time domain.

$S_{\mathrm{bb}}^{\mathrm{norm}}(\lambda, \mu)$ to the best matching frequency regions in $\hat{U}_{\mathrm{eb}}(\lambda, \mu)$ (e.g., based on a correlation measure). The method is, e.g., applied in [Tammi et al. 2009, Laaksonen et al. 2010]. With this approach, the baseband signal effectively serves as an *adaptive codebook* for the frequency subbands of the extension band excitation. The mapping of source indices μ_{s} within $S_{\mathrm{bb}}^{\mathrm{norm}}(\lambda, \mu)$ to the target indices μ_{t} has to be transmitted as side information.

Time Domain Implementation

As a new proposal, the blind replication method of (2.33) can also be implemented in the time domain. In this case, as a replacement for the spectral normalization according to (2.32), linear prediction according to (2.13) can be applied, i.e., the LP residual of the baseband signal could directly be used as excitation $\hat{u}_{\mathrm{eb}}(k)$ for the extension band synthesis. However, considering the example from above, where the 1 – 4 kHz range of the baseband signal shall be used as the excitation signal, additional processing has to be carried out. The respective block diagram is shown in Figure 2.9. To illustrate the functionality of the proposed system for spectral replication, example spectra of intermediate signals are shown in the figure. With this system, in effect, the normalized LP residual of the baseband signal $u_{\mathrm{bb}}^{\mathrm{norm}}(k)$ is suitably modulated in an upsampled domain by multiplication with the term $\cos(\omega_0\, k')$. If, as for the example spectrum plots in Figure 2.9, the modulation frequency is set to $\omega_0 = \frac{5\,\pi}{8}$, the 0 – 1 kHz range of the original input signal can be eliminated by the second low-pass filter. In case the baseband signal only covers the *telephone frequency band* with its cutoff frequency of approximately 3.5 kHz, a modulation frequency of $\omega_0 = \frac{9\,\pi}{16}$ must be chosen so that the 0.5 – 3.5 kHz spectrum is selected as extension band excitation. The final "compression" module is used to eliminate strong peaks in the time domain which may influence subsequent signal synthesis blocks (e.g., LPC synthesis filtering, see Figure 2.6).

The two low pass filters of the replication system can be realized as IIR filters with low algorithmic delay and complexity since mild phase distortions are not detrimental in the context of bandwidth extension.

2.5.2 Parameter-Driven Synthesis

As an alternative to replication based regeneration of the spectral details in the extension band, a full synthesis of these signals can be pursued. The simplest approach is to use a pseudo-random noise signal. However, tonal signal segments such as voiced speech can not be accurately modeled with a pure noise excitation.

Therefore, typical parameter-driven excitation generation methods make an explicit distinction between noisy and tonal contributions. Often, the additional assumption is made that the tonal components are purely *harmonic*. Such a mixture of noisy and harmonic signal components is in line with the commonly assumed model of *speech production*, e.g., [Vary & Martin 2006]. Nevertheless, as shown in Section 3.3.3, such a simplified model (i.e., disregarding the non-harmonic, but still tonal signal components) can also be successfully applied to more generic audio material (e.g., music). Of course, there are more sophisticated methods that allow to add or remove individual sinusoids from the generated signal. This is, e.g., applied in [Dietz et al. 2002].

For a fully synthetic excitation signal, comprising noise and harmonic components, the following parameters are useful:

- Tonality — The weighting of tonal components and noise must be deduced from a measured tonality of either the received baseband signal (no side information transmitted) or of the original extension band tonality (parameter is sent as side information).

- Pitch period — The pitch period can be estimated from the baseband signal (no side information), which is fully sufficient for speech bandwidth extension, or it can be determined from the original extension band signal (side information is transmitted).

- Pulse shape — If the excitation synthesis is performed in the time domain, a well-designed shape of the pitch pulses can help to replicate the original spectral details more accurately compared to unit pulses. Unit pulses usually result in a rather sharp sounding excitation signal. The pulse shape design can be fixed (no side information) or it can be determined based on the original signal (with side information).

The implementation of the signal generator and analysis modules can either be carried out in the frequency domain or in the time domain, as required by the given processing framework (transform codec or time domain bandwidth extension). More details and two concrete implementation examples for both possibilities are described in Section 3.2.3 (time domain implementation for wideband speech signals) and in Section 3.3.3 (frequency domain implementation for superwideband speech and audio signals), respectively.

2.5.3 Hybrid Approaches

To better accommodate the varied spectral structure of generic audio signals such as music, hybrid algorithms can be used for the regeneration of spectral details in the extension band. These algorithms switch between two (or more) regeneration methods depending on the characteristics of the current signal segment.

For example, in segments where there is sufficient similarity between the spectral details of the baseband and of the extension band signal, spectral replication techniques (as described in Section 2.5.1) are well applicable. In other cases, especially for signals that have a distinctive harmonic structure in the high band, artificial signal synthesis (Section 2.5.2) is better suited to achieve a high quality.

Hybrid approaches for the regeneration of spectral details are only reasonably applicable if the transmission of side information is allowed. At least a binary flag is required to indicate the signal regeneration method to be applied (e.g., spectral replication or pure synthesis). Naturally, a concrete hybrid method must define specific criteria to decide upon the regeneration mode to be used at the decoder.

Hybrid signal generators that switch between several modes of operation are in fact applied in several speech and audio codecs, e.g., [Laaksonen et al. 2010, Geiser et al. 2009]. In particular, the excitation generator of [Geiser et al. 2009, Geiser, Krüger & Vary 2010] which is based on switched spectral replication and harmonic synthesis is described in Section 3.3.3.

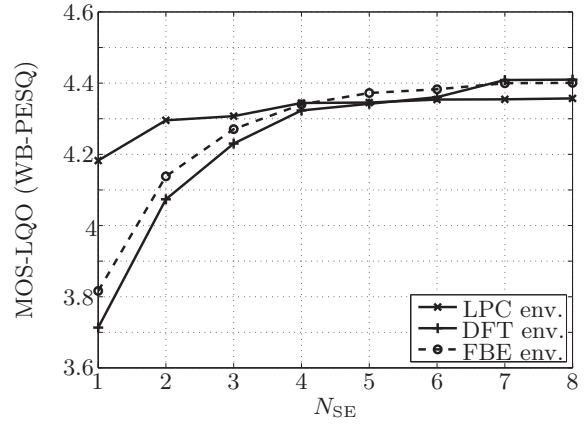
2.6 Performance of Different Parameter Sets

The different algorithmic components for bandwidth extension that have been described in the previous sections shall now be evaluated and compared. The available degrees of freedom (in particular the temporal and spectral resolution) shall be investigated for the narrowband to wideband as well as for the wideband to super-wideband bandwidth extension scenarios, both for speech and music input signals. The quality assessment is conducted with two objective audio quality measures:

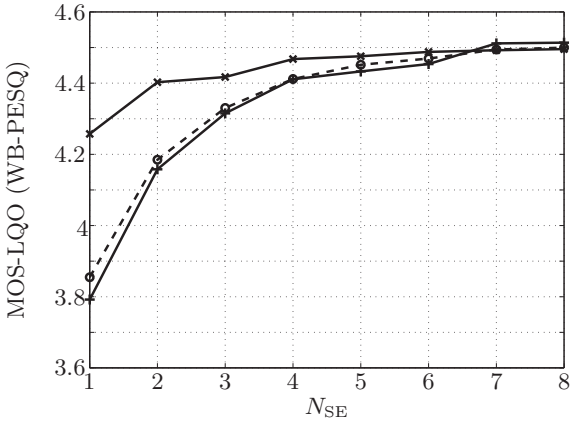
- The wideband PESQ tool [ITU-T 2005] is used to assess the quality of processed wideband speech signals (16 kHz sampling rate). The (mapped) WB-PESQ scale, given as MOS-LQO[5], ranges from approx. 1.0 for the worst quality up to 4.6 for the best quality.

- The PEAQ tool [ITU-R 1998] is used to assess the quality of processed audio signals with 16 kHz and 32 kHz sampling rate (wideband and super-wideband). The PEAQ scale, given as ODG[6], ranges from -4 for the worst quality up to 0 for the best quality.

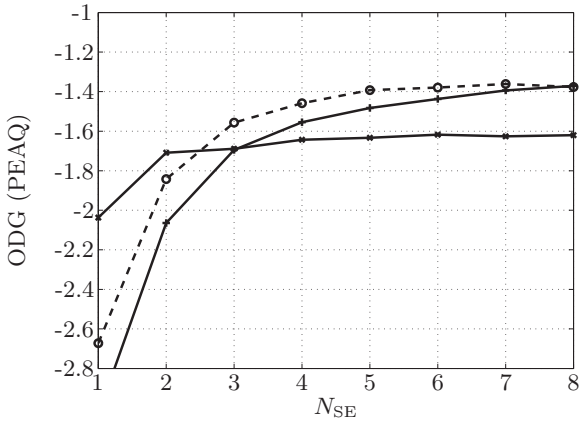
[5] MOS-LQO: Mean Opinion Score, Listening Quality Objective
[6] ODG: Objective Difference Grade



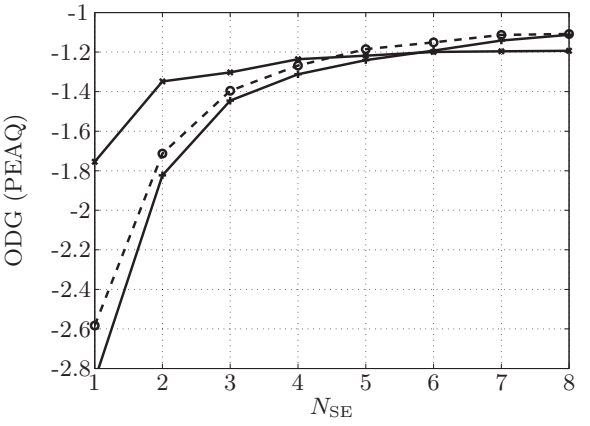
(a) $f_{\mathrm{s}}' = 16\,\mathrm{kHz},\ N_{\mathrm{TE}} = 1$



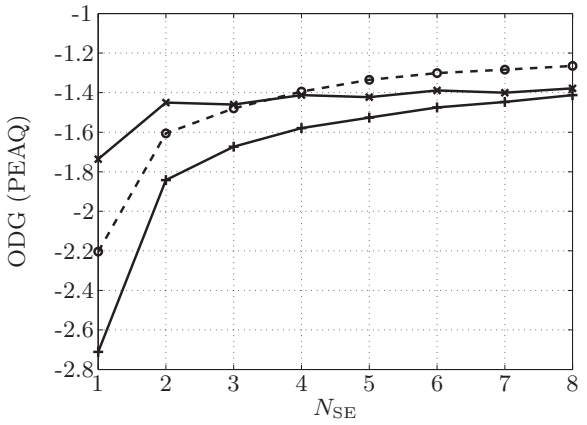
(b) $f_{\mathrm{s}}' = 16\,\mathrm{kHz},\ N_{\mathrm{TE}} = 2$



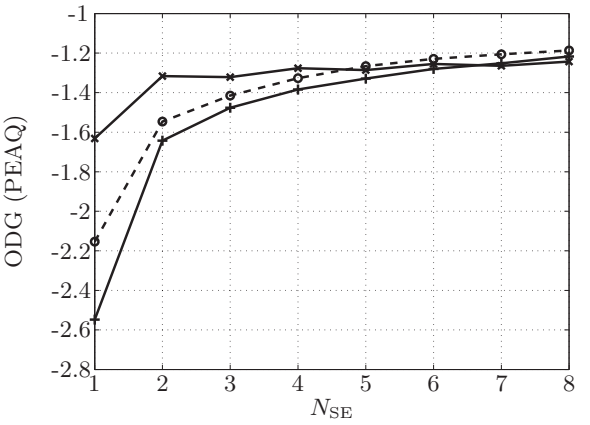
(c) $f_{\mathrm{s}}' = 16\,\mathrm{kHz},\ N_{\mathrm{TE}} = 1$



(d) $f_{\mathrm{s}}' = 16\,\mathrm{kHz},\ N_{\mathrm{TE}} = 2$

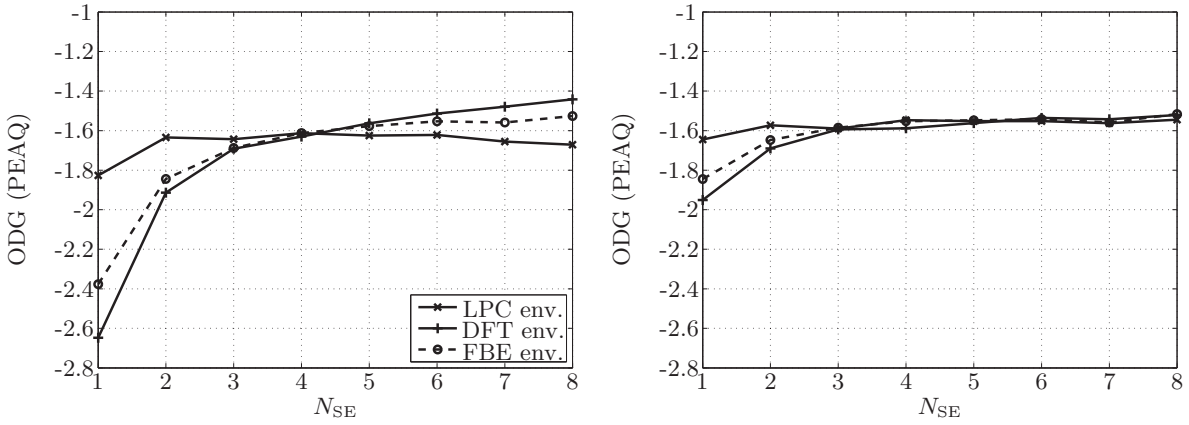


(e) $f_{\mathrm{s}}' = 32\,\mathrm{kHz},\ N_{\mathrm{TE}} = 1$



(f) $f_{\mathrm{s}}' = 32\,\mathrm{kHz},\ N_{\mathrm{TE}} = 2$

Figure 2.10: Results for wideband ($f_{\mathrm{s}}' = 16\,\mathrm{kHz}$) and super-wideband ($f_{\mathrm{s}}' = 32\,\mathrm{kHz}$) speech signals, replication excitation.



(a) $f_{\mathrm{s}}' = 16\,\mathrm{kHz}$, $N_{\mathrm{TE}} = 2$ (b) $f_{\mathrm{s}}' = 32\,\mathrm{kHz}$, $N_{\mathrm{TE}} = 2$

Figure 2.11: Exemplary results for super-wideband ($f_{\mathrm{s}}' = 32\,\mathrm{kHz}$) music signals, replication excitation.

For the present investigation, the *parallel* analysis structure of Figure 2.3(b) has been used. The frame shift has been set to 20 ms while the window size for spectral analysis and synthesis has been set to 25 ms, i.e., there is a relative window overlap of 20%. This setup is fully compatible with conversational codecs, e.g., the AMR or AMR-WB codecs as deployed in 3GPP cellular networks. For the parametrization of the extension band signals, several options have been considered: The *temporal envelope* is always represented by N_{TE} subframe gains (cf. Section 2.2) while, for the *spectral envelope*, both the autoregressive representation with an LPC filter order of N_{SE} (cf. Section 2.3) as well as the identical number of DFT domain subband gains (cf. Section 2.4) are investigated. Furthermore, for the latter case, both DFT domain synthesis of the spectral envelope (cf. Section 2.4.3) as well as time domain synthesis with a filterbank equalizer (cf. Section 2.4.4) are considered. The decoder side regeneration of the *spectral details* in the extension band is achieved by the spectral replication approach as described in Section 2.5.1, either in the time or in the frequency domain, as appropriate. The parameters have not been quantized.

The test results for bandwidth extension of speech signals are shown in Figure 2.10. The test set comprised approximately 100 samples in English language from the NTT corpus [NTT 1994]; the graphs show the averaged quality scores. In the experiments, the number of spectral parameters N_{SE} has been varied from 1 to 8, while the number of temporal parameters N_{SE} has been limited to 1 (no subframes) and 2 (10 ms subframe length). With a higher temporal resolution, only relatively small quality improvements could be obtained. A possible explanation is that the spectral replication method, which is used to regenerate the spectral details, partially preserves the temporal structure of the baseband signal and thus less gain correction is required. Nevertheless, a higher number of temporal pa-

rameters, i.e., a smaller subframe length might still be required for an adequate representation of heavily transient signals.

From Figure 2.10, it can be concluded that excellent speech quality is achievable in all investigated setups, i.e., for all envelope parameter sets at both sampling rates (16 kHz and 32 kHz). Thereby, the LPC envelope parametrization has a certain advantage because less parameters are needed to achieve the same quality level. The reason is that the frequency responses of the all-pole AR synthesis filters allow for a much more flexible allocation of the available spectral resolution. With a higher number of spectral parameters, however, the subband approaches (DFT and FBE envelope synthesis) are able to close the quality gap (or even surpass the quality of the LPC envelope synthesis).

To provide a comparison, also the *serial* signal analysis approach of Figure 2.3(a) has been tested with an exemplary setup ($f_\mathrm{s}' = 16\,\mathrm{kHz}$, LPC envelope, $N_\mathrm{SE} = 4$, $N_\mathrm{TE} = 2$). The respective WB-PESQ score reveals no measurable quality loss compared to the parallel system, which can be explained by the strict low-pass character (limited leakage) of the employed temporal gain function, see Figure 2.4(c). Consequently, a given parallel signal analysis can be easily replaced by a serial system if required by the concrete application.

In another experiment, the applicability of the proposed systems for bandwidth extension of *audio* signals (music) was assessed. The test material was taken from the EBU SQAM corpus [EBU 1988] (samples 55 – 70). The averaged test results (PEAQ scores) are shown in Figure 2.11. Obviously, there is less dependency on the number of spectral parameters than for speech input signals. However, the overall achievable quality level is lower than in the speech case. The main reason for this is that an adequate regeneration of the *spectral details* is difficult to achieve with a plain spectral replication method, especially for samples that include tonal components in the extension band. In these cases, more sophisticated methods are required to concisely regenerate the spectral details. An example algorithm that is suited for music signals is presented in the following chapter (Section 3.3). Nevertheless, also with the plain spectral replication method as investigated here, an acceptable quality level can be obtained for many audio samples, in particular for bandwidth extension towards super-wideband frequencies.

2.7 Discussion

The compact parameter sets that have been proposed in this chapter are suited to resynthesize the (wideband and super-wideband) extension band of speech signals with excellent quality. A reasonable quality level can also be maintained for more generic audio signals, e.g., music.

An important application for the newly introduced parameter sets is *embedded coding* where *quantized* parameters are added to the bitstream of the core codec. This corresponds to Option 1 in Figure 2.1. Two concrete novel algorithms that have been developed in the context of international standardization projects are

presented in the following chapter. As a second important application, parameter *estimation* can be conducted, facilitating an *artificial* bandwidth extension where only the decoder side of the communication system needs to be modified. This is addressed in Chapter 4.

The evaluation in Section 2.6 has shown that an excellent quality level is in fact achievable with *all* of the proposed algorithms and methods. Therefore, if the reproduced audio bandwidth of a *given* communication system shall be enhanced, the particular bandwidth extension algorithm can be chosen which best matches the system requirements and the given baseband codec. A few example application scenarios shall be outlined here:

- In the basic scenario where the reproduced bandwidth of a *time domain* codec (e.g., CELP) shall be enhanced with bandwidth extension techniques, a full time-domain solution should be applied so that the algorithmic delay is not increased too much. For example, the autoregressive spectral envelope of Section 2.3 could be used in conjunction with the novel temporal envelope control mechanism of Section 2.2. The spectral details could be resynthesized at the decoder side with the help of the new time domain spectral replication method of Section 2.5.1.

- If the reproduced bandwidth of a given *transform* codec shall be enhanced, a mixed frequency *and* time domain bandwidth extension method can be applied. The respective algorithms are based on frequency domain spectral replication and on the new temporal gain manipulation method. The *spectral* envelope synthesis is carried out in the frequency domain, cf. Section 2.4.3. Typically, the overall algorithmic delay is not increased significantly in such a scenario. As an additional benefit of the transform domain solution, additional enhancement layers can be added to the codec to (selectively) refine the synthetic extension band signal with quantized spectral coefficients. A real-world example for such a codec is presented in the following chapter.

- The third application scenario is a variant of the transform domain solution. In a hybrid hierarchical codec that comprises a time-domain core codec and one or more frequency domain extension layers, the *encoder part* of the bandwidth extension algorithm can be realized in the frequency domain. At the *decoder*, the spectral envelope can be resynthesized *either* in the frequency domain (as above) *or* in the time domain, e.g., by using the concept of the filterbank equalizer (Section 2.4.4) which is a new proposal in this thesis. Thereby, the same parameter set can be used for both approaches. The advantage is improved flexibility: The time domain decoder can reduce the algorithmic delay because the decoder side overlap-add operation (which is needed for transform coding with half-overlapped windows) can be omitted. In contrast, the frequency domain realization of the decoder algorithms facilitates a refinement of spectral coefficients by quantized extension layers.

Chapter 3

Bandwidth Extension for Embedded Speech and Audio Coding

The signal processing algorithms of Chapter 2 can be directly applied to upgrade the audio bandwidth of a given narrowband (or wideband) codec. To accomplish the integration with a codec, appropriate (vector) quantization techniques need to be applied to the bandwidth extension parameters. The resulting bits are then transmitted within an "add-on" bitstream layer. Effectively, such a layered bitstream arrangement forms a special case of the *embedded coding* paradigm, cf. [Geiser, Ragot & Taddei 2008].

In this chapter, first, a brief introduction to the underlying principles and ideas of embedded speech and audio coding is provided (Section 3.1). Then, two new bandwidth extension algorithms are described, both of which have been developed in the context of international standardization projects. The first approach, as described in Section 3.2, focuses on a time-domain solution for narrowband to wideband extension of *speech* signals. This algorithm, which is referred to as "Time Domain Bandwidth Extension" (TDBWE), was standardized in 2006 as a part of ITU-T Rec. G.729.1 [ITU-T 2006, Ragot et al. 2007]. The second method (Section 3.3) aims at a *frequency domain* extension of speech *and* music signals from wideband towards the super-wideband bandwidth. It has been included in a candidate codec for the super-wideband extensions of ITU-T Rec. G.729.1 and G.718 [ITU-T 2008*a*, Vaillancourt et al. 2008]. The proposed super-wideband codec fulfills all ITU-T requirements for mono signals.

To allow a conceptual comparison of the proposed techniques with the published state-of-the-art, the chapter concludes with a survey of related algorithms that have been proposed in the scientific literature or that have been developed in the course of other standardization projects (Section 3.4).

3.1 Embedded Speech and Audio Coding

The basic concept of embedded speech and audio coding is illustrated with the example in Figure 3.1. The encoder produces a bitstream with a layered structure, i.e., one *core layer* and, in the example, two *enhancement layers* that are stacked on top of each other. Thereby, the number of bitstream layers and the respective bit rate increments between the layers define the so-called coding *granularity*.

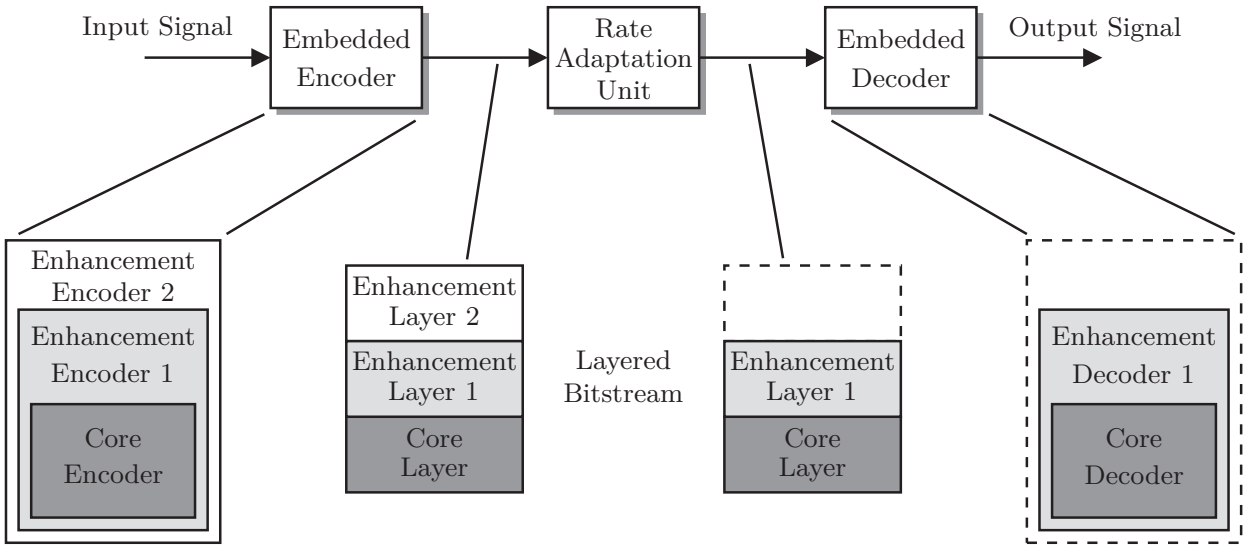


Figure 3.1: Example transmission system based on an embedded codec.

A layered bitstream structure is also *hierarchical*[1] in the sense that a given bitstream layer can only be decoded if the lower layers have been received as well. In contrast to conventional multi-mode speech codecs such as the Adaptive Multi Rate (AMR) [ETSI 2000, Ekudden et al. 1999] or Adaptive Multi Rate Wideband (AMR-WB) [ETSI 2001*b*, Bessette et al. 2002] codecs, the bit rate adaptation for an *embedded* codec is performed by simple bitstream "truncation," i.e., by simply dropping one or more layers from the hierarchical bitstream. This is achieved with a simple "rate adaptation unit" which adapts the bit rate, e.g., to current network traffic conditions or to the receiver capabilities. Note that this operation can be performed anywhere in the network without requiring a dedicated feedback channel to the encoder. The decoding of a layered bitstream is then achieved with nested or *embedded* algorithms as shown in the figure. If only the core layer is received, a decoded signal with a *basic* quality can be reproduced. As soon as enhancement layers are received, the decoder produces a signal of *enhanced quality*.

One of the main motivations for embedded coding, apart from the elegant way to adapt the bit rate, is the possibility to upgrade existing communication systems without compromising the interoperability with existing infrastructure and end user terminals. The core codec is then, typically, a widely deployed narrowband (or wideband) speech codec. In this application, the concept of enhancement layers proves to be particularly versatile because multiple functionalities can be added to a given codec, e.g., audio bandwidth extension, audio quality improvement (also called signal-to-noise ratio (SNR) scalability), or a mono to stereo extension.

[1]Note that the terminology in embedded coding is not consistent. The literature alternatively refers to this concept as *embedded, hierarchical, scalable, progressive, multi-resolution, successively refinable, or bit-droppable*. Hereafter, the terms "embedded" and "hierarchical" will be primarily used.

Embedded coding can in principle be realized in two different ways:

1. A certain codec parameter or a parameter vector may be quantized in a *hierarchical* fashion. This means that the quantized representation of the parameter can be reconstructed with *different resolutions* depending on the amount of bits received. Such a property is achieved with so-called *hierarchical (vector-)quantization techniques*, e.g., [Erdmann 2005]. This is particularly interesting if a successive refinement of the quantizer resolution, e.g., in transform codecs, is desired.

2. Alternatively, the encoder does not focus on the refinement of *existing* codec parameters, but instead adds *new* parameters to the bitstream. After rate adaptation, the decoder may only receive a *part* of the quantized parameters. Without the additional parameters, it can produce an output signal of intermediate quality. This approach can be termed "parameter dropping."

The two bandwidth extension algorithms to be discussed in this chapter fall in the latter category as an entire set of parameters is quantized and appended to the bitstream of the respective core codec. A more comprehensive overview of embedded speech and audio coding is provided in [Geiser, Ragot & Taddei 2008]. Moreover, several recent embedded codec standards that define a bitstream layer for bandwidth extension are briefly summarized in Section 3.4.

3.2 Time Domain Bandwidth Extension (TDBWE)

The TDBWE algorithm has been designed to extend the bandwidth of narrowband CELP codecs such as ITU-T G.729 or 3GPP EFR towards the wideband frequency range. The basic algorithmic concepts have been initially published in [Jax et al. 2006*a*]. This version of the algorithm constituted a part of a codec proposal by Siemens AG (Germany), Matsushita Electric Industrial Co., Ltd. (Japan), and Mindspeed Technologies, Inc. (USA). It has been submitted for the qualification phase of the ITU-T G.729.1 standardization project, see [Geiser et al. 2006]. The bit rate related to the bandwidth extension module was 2 kbit/s in this proposal. In the following, the *standardized* version [Geiser et al. 2007a] which is now a part of ITU-T Rec. G.729.1, will be summarized. Here, the bit rate for the TDBWE enhancement layer could be lowered to 1.65 kbit/s.

Section 3.2.1 defines the TDBWE parameters. Their *quantization* with a sum bit rate of 1.65 kbit/s is summarized in Section 3.2.2. To complete the technical description, the TDBWE *synthesis* algorithm is detailed in Section 3.2.3. The integration of the entire algorithm in the framework of the ITU-T G.729.1 codec is explained in Section 3.2.4. Finally, a comprehensive evaluation and a discussion is provided (Sections 3.2.5 and 3.2.6).

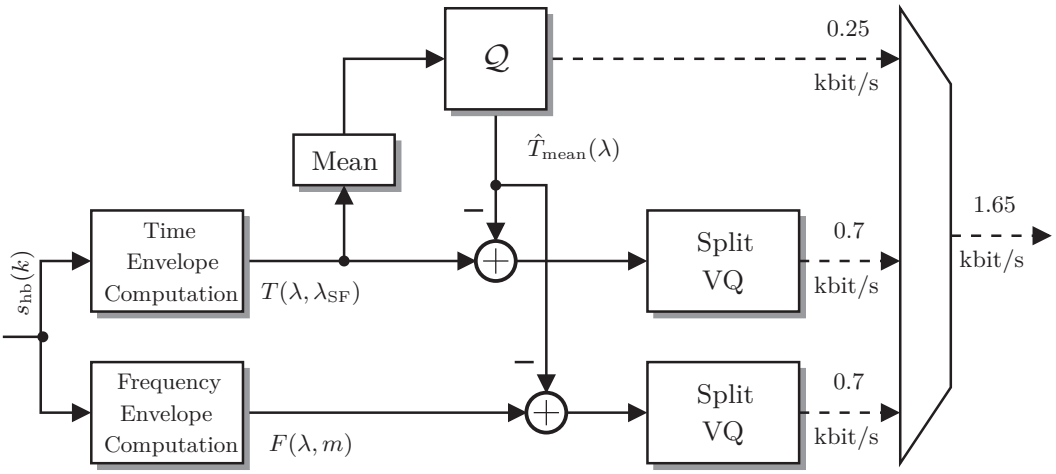


Figure 3.2: TDBWE encoder: Parameter extraction and quantization.

3.2.1 Parameter Set for Bandwidth Extension

The TDBWE encoder, depicted in Figure 3.2, operates on 20 ms frames of the downsampled ($f_{\text{s}} = 8\,\text{kHz}$) and pre-processed (lowpass with $f_{\text{c}} = 3\,\text{kHz}$) high band signal $s_{\text{hb}}(k)$ that has been obtained with a half-band QMF analysis filterbank, see Section 2.1.2. This signal is used in the following as a special case of the generic extension band signal $s_{\text{eb}}(k)$ from the previous chapter. Note that, owing to the downsampling and pre-filtering, the high band signal $s_{\text{hb}}(k)$ comprises frequencies between 0 and 3 kHz. Naturally, these frequencies describe the original high band frequency range of 4 – 7 kHz.

In the TDBWE algorithm, the *parallel* signal analysis approach of Figure 2.3(b) is used. The respective parameter set encompasses a *temporal envelope* and a *spectral envelope*, here also denoted *time* and *frequency* envelopes. The determination of these parameters which are variants of the temporal subframe gains of Section 2.2.1 and of the spectral subband gains of Section 2.4.2 is summarized below.

Temporal Envelope

The TDBWE algorithm uses a straight-forward representation of the temporal envelope in terms of subframe gains as introduced in Section 2.2.1. These gains are used to determine a temporal gain function of *fixed* resolution. The 20 ms ($L = 160$) input frame of the signal $s_{\text{hb}}(k)$ with frame index λ is subdivided into $N_{\text{TE}} = 16$ segments of length 1.25 ms each, i.e., each segment comprises $L_{\text{SF}} = 10$ samples. In contrast to (2.5), the time envelope parameters $T(\lambda, \lambda_{\text{SF}})$ with $\lambda_{\text{SF}} \in \{0, \ldots, N_{\text{TE}} - 1\}$ are now computed as *logarithmic* subframe gains

$$T(\lambda, \lambda_{\text{SF}}) = \frac{1}{2}\,\text{ld}\,\frac{1}{L_{\text{SF}}} \sum_{k=0}^{L_{\text{SF}}-1} s_{\text{hb}}^2(\lambda L + \lambda_{\text{SF}} L_{\text{SF}} + k). \tag{3.1}$$

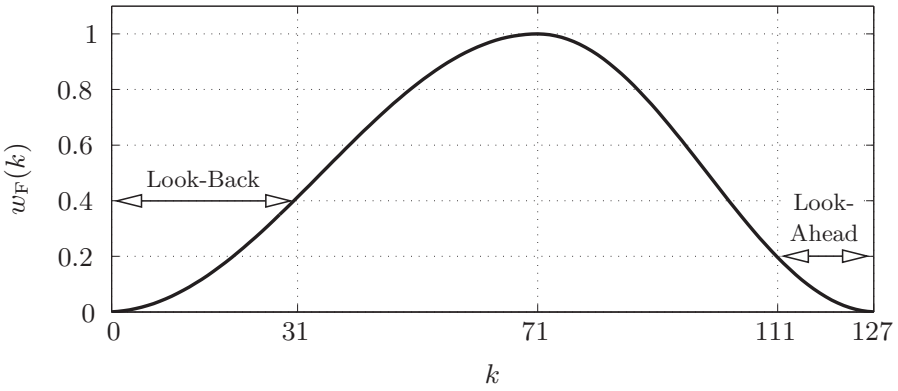


Figure 3.3: Window function $w_{\mathrm{F}}(k)$ for spectral envelope computation.

The logarithm operation is applied to facilitate a (perceptually advantageous) logarithmic quantization with a simple uniform quantizer, cf. Section 3.2.2. The *binary* logarithm $\mathrm{ld}\, x \doteq \log x / \log 2$ has been chosen to ease an implementation in fixed point arithmetic.

Spectral Envelope

The spectral envelope of the TDBWE parameter set is represented in terms of $N_{\mathrm{SE}} = 12$ DFT domain subband gains for each signal frame, similar to the description of Section 2.4. For the computation of the respective parameters $F(\lambda, m)$ with $m \in \{0, \ldots, 11\}$, the signal $s_{\mathrm{hb}}(k)$ is windowed by a slightly asymmetric analysis window $w_{\mathrm{F}}(k)$. This window, as shown in Figure 3.3, is 128 taps (16 ms) long and is constructed from the rising slope of a 144-tap Hann window, followed by the falling slope of a 113-tap Hann window:

$$w_{\mathrm{F}}(k) = \begin{cases} \frac{1}{2} - \frac{\cos\left(\frac{2\pi(k+1)}{144}\right)}{2}, & k \in \{0, \ldots, 71\} \\ \frac{1}{2} - \frac{\cos\left(\frac{2\pi(k-14.5)}{113}\right)}{2}, & k \in \{72, \ldots, 127\}\,. \end{cases} \tag{3.2}$$

The window is constructed such that the spectral envelope computation has a look-ahead of 16 samples (or 2 ms) and a look-back of 32 samples (or 4 ms).

As a special characteristic in the TDBWE algorithm, the maximum of $w_{\mathrm{F}}(k)$ is centered on the *second* half of the current 20 ms frame. The windowed signal for frame λ with $k \in \{0, \ldots, 127\}$ is thus given by

$$s^{\mathrm{w}}_{\mathrm{hb}}(\lambda, k) = s_{\mathrm{hb}}(\lambda L + k + 32) \cdot w_{\mathrm{F}}(k). \tag{3.3}$$

The envelope parameters for the *first part* of the 20 ms frame are in fact *not* computed. Instead, they are *interpolated* at the decoder side between the transmitted parameters from the current and from the previous frame, cf. (3.14) in Section 3.2.3.

As in Section 2.4, the windowed signal $s^{\mathrm{w}}_{\mathrm{hb}}(k)$ is transformed into the frequency domain. However, to reduce the computational load of the parameter computation,

the full Discrete Fourier Transform (DFT) of length 128 is replaced by a DFT of length 64 and a preceding polyphase network, cf. [Vary & Heute 1980]. The modified DFT computation, including the polyphase network, can be written as

$$S_{\mathrm{hb}}^{\mathrm{DFT}}(\lambda,\mu) = \frac{1}{64}\sum_{k=0}^{63}\left(s_{\mathrm{hb}}^{\mathrm{w}}(k) + s_{\mathrm{hb}}^{\mathrm{w}}(k+64)\right)\cdot \mathrm{e}^{-\jmath\frac{2\pi\mu k}{64}}, \tag{3.4}$$

where $\mu \in \{0,\ldots,63\}$. With this DFT structure, effectively, the *even bins* of the full length 128-tap DFT can be computed, i.e., a higher frequency selectivity is achieved than with a plain DFT of length 64. In the actual implementation, (3.4) is realized with a radix-2 FFT algorithm.

Finally, the frequency envelope parameter set is, similar to (2.21), calculated as logarithmic weighted subband gains for $N_{\mathrm{SE}} = 12$ overlapping, evenly spaced, and equally wide DFT domain subbands with index $m \in \{0,\ldots,N_{\mathrm{SE}}-1\}$

$$F(\lambda,m) = \frac{1}{2}\,\mathrm{ld}\sum_{\mu=0}^{M_{\mathrm{SB}}-1} W(\mu)\cdot\left|S_{\mathrm{hb}}^{\mathrm{DFT}}(\lambda,\mu+mM_{\mathrm{s}})\right|^2, \tag{3.5}$$

where $M_{\mathrm{s}} = 2$ and $M_{\mathrm{SB}} = 3$, i.e., the m-th subband starts at the DFT bin with index $m\cdot M_{\mathrm{s}}$ and spans a bandwidth of M_{SB} DFT bins. This corresponds to a physical subband division with a bandwidth of $\Delta f_m = 375\,\mathrm{Hz}$ for each subband (except the first one, which amounts to $\Delta f_0 = 312.5\,\mathrm{Hz}$). The frequency bins with indices 25 – 31 are not considered since they represent frequencies above 3 kHz. The frequency domain weighting window $W(\mu)$ in (3.5), is given as

$$W(\mu) = \begin{cases} 0.5, & \mu = 0 \\ 1.0, & \mu = 1 \\ 0.5, & \mu = 2 \end{cases}. \tag{3.6}$$

3.2.2 Quantization

The TDBWE parameter set (consisting of the temporal envelope parameters $T(\lambda,\lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{0,\ldots,N_{\mathrm{TE}}-1\}$ and of the spectral envelope parameters $F(\lambda,m)$ with $m \in \{0,\ldots,N_{\mathrm{SE}}-1\}$) is quantized using *mean-removed split vector quantization* (cf. Figure 3.2). Therefore, the *mean time envelope* per frame

$$T_{\mathrm{mean}}(\lambda) = \frac{1}{N_{\mathrm{TE}}}\sum_{\lambda_{\mathrm{SF}}=0}^{N_{\mathrm{TE}}-1} T(\lambda,\lambda_{\mathrm{SF}}). \tag{3.7}$$

is quantized with a wordlength of 5 bits using uniform 3 dB steps in the logarithmic domain. The quantized value $\hat{T}_{\mathrm{mean}}(\lambda)$ is then subtracted from the entire parameter set, i.e.,

$$T'(\lambda,\lambda_{\mathrm{SF}}) = T(\lambda,\lambda_{\mathrm{SF}}) - \hat{T}_{\mathrm{mean}}(\lambda) \quad \text{and} \quad F'(\lambda,m) = F(\lambda,m) - \hat{T}_{\mathrm{mean}}(\lambda). \tag{3.8}$$

Table 3.1: Bit allocation for TDBWE parameter quantization.

Parameter / Vector	Dimension	No. allocated bits (per 20 ms)
$T_{\mathrm{mean}}(\lambda)$	1	5
$\mathbf{T'_1}(\lambda)$	8	7
$\mathbf{T'_2}(\lambda)$	8	7
$\mathbf{F'_1}(\lambda)$	4	5
$\mathbf{F'_2}(\lambda)$	4	5
$\mathbf{F'_3}(\lambda)$	4	4
$\sum$	29	$33 \triangleq 1.65\,\mathrm{kbit/s}$

By this subtraction, the obtained values become independent from the overall signal level. Note that, due to the logarithm operation, the parameter $\hat{T}_{\mathrm{mean}}(\lambda)$ in fact corresponds to a *geometric* mean of the subframe gains. However, no significant quality difference could be observed when using the arithmetic mean instead. The mean removed time envelope parameter set is gathered in two vectors of dimension eight ($\mathbf{T'_1}(\lambda)$ and $\mathbf{T'_2}(\lambda)$) whereas the frequency envelope parameter set forms three vectors of dimension four ($\mathbf{F'_1}(\lambda)$ – $\mathbf{F'_3}(\lambda)$). Finally, vector quantization based on pre-trained quantization tables (codebooks) with the bit allocation from Table 3.1 is applied. The individual codebooks for the five subvectors have been constructed with the well-known LBG algorithm [Linde et al. 1980]. Yet, to maintain a certain pairwise distance between the centroids, the LBG-codebook is requantized using a rectangular grid with a step size of 6 dB in the logarithmic domain. In the TDBWE decoder, the quantized temporal and spectral envelopes are then computed by adding the decoded mean value $\hat{T}_{\mathrm{mean}}$:

$$\hat{T}(\lambda, \lambda_{\mathrm{SF}}) = \hat{T}'(\lambda, \lambda_{\mathrm{SF}}) + \hat{T}_{\mathrm{mean}}(\lambda), \quad \text{and} \quad \hat{F}(\lambda, m) = \hat{F}'(\lambda, m) + \hat{T}_{\mathrm{mean}}(\lambda). \tag{3.9}$$

The presented quantization scheme with its total bit rate of $33\,\mathrm{bit}/20\,\mathrm{ms} = 1.65\,\mathrm{kbit/s}$ is evaluated in Section 3.2.5 (Table 3.2) by comparing the obtained speech quality with that of an unquantized TDBWE parameter set.

3.2.3 Synthesis

The TDBWE decoder, a block diagram is depicted in Figure 3.4, uses the decoded parameters $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ and $\hat{F}(\lambda, m)$ to appropriately shape an artificially generated excitation signal $\hat{u}_{\mathrm{hb}}(k)$. In contrast to the processing order of Figure 2.3(c), the TDBWE algorithm restores the *temporal* envelope *before* the *spectral* envelope. In addition, an adaptive *post-processing* procedure, i.e., amplitude compression, is implemented. These algorithmic modules are detailed in the following.

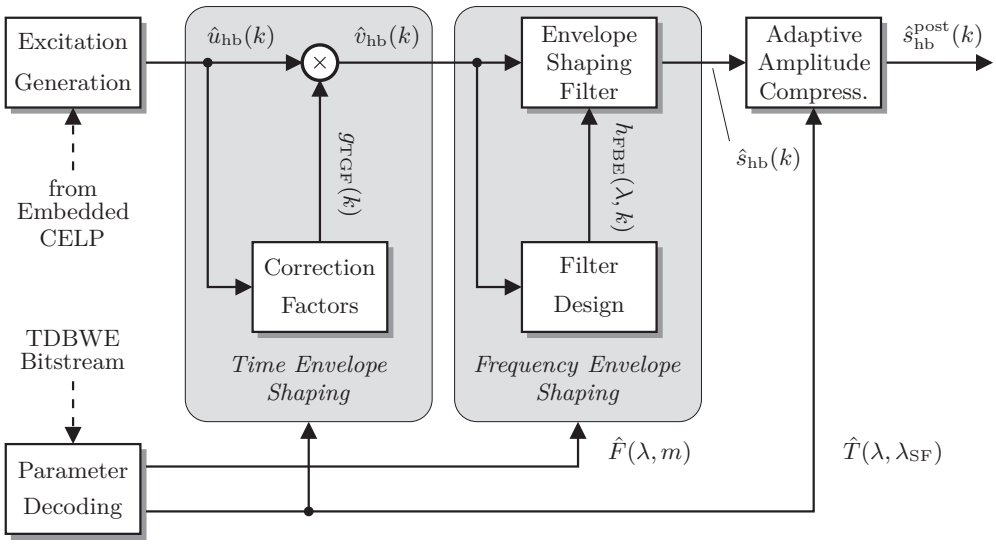


Figure 3.4: Block diagram of the TDBWE decoder.

Excitation Signal Generation

The excitation signal generator of the TDBWE algorithm is a concrete realization of the parameter-driven excitation synthesis method as introduced in Section 2.5.2. It is suited to synthetically regenerate the spectral details in the high frequency band of wideband speech signals (4 – 7 kHz). The algorithm operates entirely in the time domain. A block diagram is shown in Figure 3.5.

The TDBWE parameter set, as described in the previous section, comprises no explicit description of the spectral details. Instead, in order to replicate the desired speech-like behavior of the high band excitation signal, several parameters of the underlying CELP core layer codec are reused. In the following, the algorithm, as shown in Figure 3.5, shall be briefly outlined. The individual algorithmic steps are discussed in much more detail in Appendix A of this thesis. Further note that (sub)frame indices are omitted in this section for notational convenience.

The TDBWE excitation signal $\hat{u}_{\mathrm{hb}}(k)$ is generated on a 5 ms subframe basis which stems from the subframe division of the narrowband core codec. It is produced as a weighted mixture of noisy (unvoiced) and periodic (voiced) components. The algorithm is structured as follows:

1. Gain estimation
   Two gains g_{v} and g_{uv} are estimated to weight the voiced and unvoiced contributions to the excitation signal $\hat{u}_{\mathrm{hb}}(k)$. The voiced gain g_{v} is essentially a post-processed version of the energy ratio of the adaptive and fixed codebook contributions of the narrowband CELP codec (cf., e.g., Section 5.4.1). The post-processing aims at a more consistent temporal evolution of the voiced gain. The unvoiced gain is then simply derived as $g_{\mathrm{uv}} = \sqrt{1 - g_{\mathrm{v}}^2}$.

2. Pitch lag post-processing
   As the voiced contribution to the excitation signal can be assumed to be a

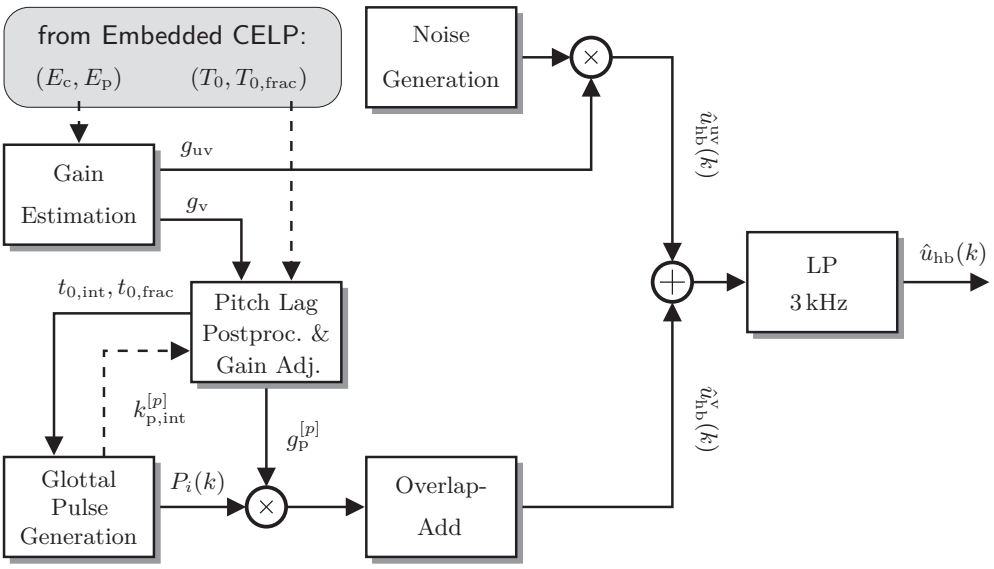


Figure 3.5: TDBWE decoder: Excitation signal generation.

harmonic continuation of the narrowband signal, the (fractional) *pitch lag* parameters (T_0 and $T_{0,\mathrm{frac}}$) of the core layer codec can be reused. However, since the criterion to determine T_0 and $T_{0,\mathrm{frac}}$ in the core codec is, typically, a maximized prediction gain, the correct harmonic pitch contour (fundamental frequency F_0) is not necessarily represented by these parameters. Therefore, the pitch post-processing procedure tries to remove typical pitch estimation errors, in particular pitch doubling errors. See Appendix A for details.

3. Production of the voiced contribution
   The voiced contribution $\hat{u}_{\mathrm{hb}}^{\mathrm{v}}(k)$ to the excitation signal is produced by an overlap-add of weighted, spectrally shaped and suitably spaced *glottal pulses*. The pulse spacing is determined by the integer portion $t_{0,\mathrm{int}}$ of the post-processed pitch lag. To accommodate a *fractional pitch resolution*, here of 1/6 of a sample, the pulse *shape* is determined by the fractional portion $t_{0,\mathrm{frac}}$ of the post-processed pitch lag. Furthermore, since the pulses also exhibit a specific *spectral* shape (a slight lowpass character), an overvoicing at high frequencies can be avoided and the natural speech characteristic is represented rather accurately.

4. Production of the unvoiced contribution
   The unvoiced contribution $\hat{u}_{\mathrm{hb}}^{\mathrm{uv}}(k)$ is produced with the weighted output of a simple pseudo random noise generator.

5. Lowpass filtering
   Finally, a 3 kHz lowpass filter is applied to the sum of the voiced and unvoiced contributions. This filter limits the frequency content of the output signal to a cutoff frequency of 7 kHz in the wideband domain, i.e., after QMF filterbank synthesis.

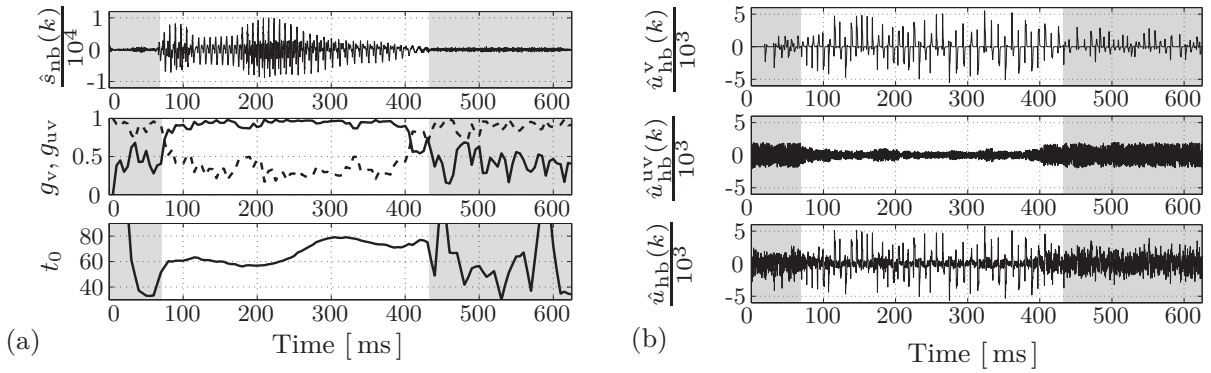


Figure 3.6: (a) Lower band speech signal $\hat{s}_{\mathrm{nb}}(k)$ and parameters for the excitation signal generation: Voiced gain g_{v} (solid line), unvoiced gain g_{uv} (dashed line), and post-processed pitch lag $t_0 \doteq t_{0,\mathrm{int}} + t_{0,\mathrm{frac}}/6$. The shown speech fragment represents an unvoiced/voiced/unvoiced transition. — (b) Example of the generated high band excitation: *Voiced* and *unvoiced contributions* as well as the final (lowpass filtered) excitation signal. The parameters for the signal fragment from (a) are used.

To illustrate the operation of the excitation generation algorithm, Figure 3.6(a) shows the parameters g_{v}, g_{uv}, and $t_0 \doteq t_{0,\mathrm{int}} + t_{0,\mathrm{frac}}/6$ which are obtained from the received low band speech signal segment $\hat{s}_{\mathrm{nb}}(k)$ shown in the example. In particular, it can be observed that the pitch contour evolves very smoothly during the voiced period. The individual contributions to the excitation signal, the production of which is based on these parameters, are visualized in Figure 3.6(b).

Temporal Envelope Shaping

The TDBWE temporal envelope shaping is realized with a temporal gain function of fixed resolution (see Section 2.2.1) based on the subframe length of 1.25 ms, i.e.,

$$\hat{v}_{\mathrm{hb}}(k) = \hat{u}_{\mathrm{hb}}(k) \cdot \hat{g}_{\mathrm{TGF}}(k). \tag{3.10}$$

To establish the temporal gain function $\hat{g}_{\mathrm{TGF}}(k)$, the received and decoded time envelope parameters $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ as well as the measured parameters $T_{\hat{u}}(\lambda, \lambda_{\mathrm{SF}})$ which are related to the excitation signal $\hat{u}_{\mathrm{hb}}(k)$ are used, whereby $\lambda_{\mathrm{SF}} \in \{0, \ldots, N_{\mathrm{TE}} - 1\}$ with $N_{\mathrm{TE}} = 16$. To determine $T_{\hat{u}}(\lambda, \lambda_{\mathrm{SF}})$, the excitation signal $\hat{u}_{\mathrm{hb}}(k)$ is segmented and analyzed in the same manner as described in Section 3.2.1 for the parameter extraction in the encoder. Then, similar to (2.8), relative gains

$$\hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}}) = 2^{\hat{T}(\lambda, \lambda_{\mathrm{SF}}) - T_{\hat{u}}(\lambda, \lambda_{\mathrm{SF}})}. \tag{3.11}$$

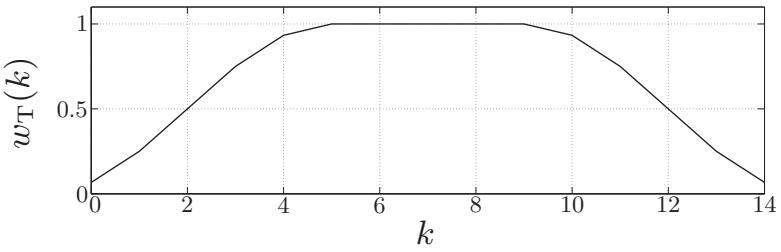


Figure 3.7: "Flat-top" Hann window for the temporal envelope shaping.

are computed. Now, for each signal segment with $L_{\mathrm{SF}} = 10$, these gain factors are interpolated using a "flat-top" Hann window

$$w_{\mathrm{T}}(k) = \begin{cases} \frac{1}{2} - \frac{\cos\left(\frac{(k+1)\cdot\pi}{L_{\mathrm{SF}}/2-1}\right)}{2}, & k \in \left\{0, \ldots, \frac{1}{2}L_{\mathrm{SF}} - 1\right\} \\ 1 \quad , & k \in \left\{\frac{1}{2}L_{\mathrm{SF}}, \ldots, L_{\mathrm{SF}} - 1\right\} \\ \frac{1}{2} - \frac{\cos\left(\frac{(k+L_{\mathrm{SF}}-1)\cdot\pi}{L_{\mathrm{SF}}/2-1}\right)}{2}, & k \in \left\{L_{\mathrm{SF}}, \ldots, \frac{3}{2}L_{\mathrm{SF}} - 1\right\}, \end{cases} \tag{3.12}$$

which is plotted in Figure 3.7. The interpolated temporal gain function $\hat{g}_{\mathrm{TGF}}(k)$ is finally computed as

$$\begin{aligned} &\hat{g}_{\mathrm{TGF}}(\lambda L + \lambda_{\mathrm{SF}} L_{\mathrm{SF}} + k) \\ &\quad = \begin{cases} w_{\mathrm{T}}(k)\, \hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}}) + w_{\mathrm{T}}(k + L_{\mathrm{SF}})\, \hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}} - 1), \\ \qquad\qquad\qquad\qquad\qquad\qquad k \in \left\{0, \ldots, \frac{1}{2}L_{\mathrm{SF}} - 1\right\} \\ w_{\mathrm{T}}(k)\, \hat{g}_{\mathrm{rel}}(\lambda, \lambda_{\mathrm{SF}}), \\ \qquad\qquad\qquad\qquad\qquad\qquad k \in \left\{\frac{1}{2}L_{\mathrm{SF}}, \ldots, L_{\mathrm{SF}} - 1\right\}, \end{cases} \end{aligned} \tag{3.13}$$

where, in analogy to (2.6), $\hat{g}_{\mathrm{rel}}(\lambda, -1) \doteq \hat{g}_{\mathrm{rel}}(\lambda - 1, N_{\mathrm{TE}} - 1)$, i.e., this gain factor is taken from the *last* 1.25 ms segment of the *preceding* frame. Compared to the gain interpolation of (2.6), only the current and the previous gain factors are used here, which avoids additional look-ahead samples. However, a slight "lag" (delay) in the temporal envelope contour is accepted. This "lag" is reduced by employing the flat-top window (3.12) with its lower overlap duration.

Spectral Envelope Shaping

The TDBWE spectral envelope shaping module is a particular realization of the cosine modulated filterbank equalizer introduced in Section 2.4.4. Here, the $N_{\mathrm{SE}} = 12$ channels of the filterbank equalizer cover the 0 – 3 kHz frequency range of the extension band signal.

The received and quantized spectral envelope parameters $\hat{F}(\lambda, m)$ with $m \in \{0, \ldots, N_{\mathrm{SE}} - 1\}$ were only computed for the second half of each 20 ms frame. The first 10 ms are instead covered by linear parameter interpolation between the

current parameter set $\hat{F}(\lambda, m)$ and the parameter set $\hat{F}(\lambda - 1, m)$ from the preceding frame:

$$\hat{F}_{\text{int}}(\lambda, m) = \frac{1}{2}\left(\hat{F}(\lambda - 1, m) + \hat{F}(\lambda, m)\right). \tag{3.14}$$

For filters with finite impulse response, this coefficient interpolation is an alternative to the filter crossfading method as introduced in Section 2.3.3. A similar method is, e.g., also applied to interpolate the LPC filter coefficients in the 3GPP AMR codec, cf. [ETSI 2000, Ekudden et al. 1999].

The temporally shaped excitation signal $\hat{v}_{\text{hb}}(k)$, see (3.10), is analyzed according to the description from (3.2) – (3.5). This is done twice per frame, i.e., for the first 10 ms ($l = 1$) as well as for the second 10 ms ($l = 2$) of the current frame. The procedure yields two observed spectral envelope parameter sets $\tilde{F}_l(\lambda, m)$ with $m \in \{0, \ldots, N_{\text{SE}}\}$ and $l \in \{1, 2\}$. Now the correction gain factors $\hat{\gamma}_{\text{rel}}^{(l)}(\lambda, m)$ per subband with index m are determined for the first ($l = 1$) and for the second ($l = 2$) half of the current frame:

$$\hat{\gamma}_{\text{rel}}^{(1)}(\lambda, m) = 2^{\hat{F}_{\text{int}}(\lambda,m) - \tilde{F}_1(\lambda,m)} \quad \text{and} \quad \hat{\gamma}_{\text{rel}}^{(2)}(\lambda, m) = 2^{\hat{F}(\lambda,m) - \tilde{F}_2(\lambda,m)}. \tag{3.15}$$

These gains are used to control the individual channels of the filterbank equalizer which are defined by their bandpass filter impulse responses $h_{\text{FBE}}^{(m)}(k)$ of length $L_{\text{FBE}} = 33$ ($m \in \{0, ..., N_{\text{SE}} - 1\}$ and $k \in \{0, \ldots, L_{\text{FBE}} - 1\}$) and by a complementary highpass contribution $h_{\text{HP}}(k)$. Thereby, $h_{\text{FBE}}^{(m)}(k)$ and $h_{\text{HP}}(k)$ constitute linear phase finite impulse response (FIR) filters with a group delay of 2 ms (16 samples) each. Note that this delay exactly matches the look-ahead which is introduced by the encoder side parameter extraction, see (3.3). The filterbank equalizer is designed such that its individual channel bandwidths match the subband division which is used in (3.5). A good compromise for the (relatively short) prototype filter of length L_{FBE} is based on the *Kaiser* window design, i.e.,

$$h_0(k) = \eta \cdot \frac{I_0\left(\beta \cdot \sqrt{1 - [(k - \alpha)/\alpha]^2}\right)}{I_0(\beta)} \tag{3.16}$$

where $\alpha = (L_{\text{FBE}} - 1)/2$. In (3.16), $I_0(\cdot)$ is the modified Bessel function of the first kind with shape parameter β which has been chosen as 4. The normalization factor η is used to achieve a unity frequency response at neutral filterbank equalizer gains. Given the prototype lowpass $h_0(k)$, the individual filterbank channels' impulse responses $h_{\text{FBE}}^{(m)}(k)$ are now derived by cosine modulations according to (2.29). A complementary highpass $h_{\text{HP}}(k)$ is defined by

$$h_{\text{HP}}(k) = \delta\left(k - \frac{L_{\text{FBE}}}{2} + 1\right) - \sum_{m=0}^{N_{\text{SE}}-1} h_{\text{FBE}}^{(m)}(k) \tag{3.17}$$

for $k \in \{0, \ldots, L_{\text{FBE}} - 1\}$. Thereby, $\delta(k_0)$ is one for $k_0 = 0$ and zero otherwise. The respective frequency responses for this filterbank design are depicted in Figure 3.8.

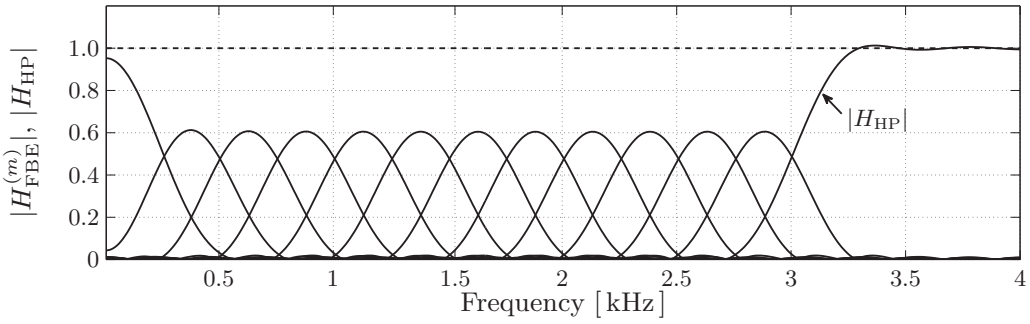


Figure 3.8: Filterbank design for the TDBWE spectral envelope shaping.

To realize the actual spectral envelope shaping, two FIR filters are constructed for each frame:

$$h_{\mathrm{FBE},l}(\lambda, k) = \sum_{m=0}^{N_{\mathrm{SE}}-1} \hat{\gamma}^{(l)}_{\mathrm{rel}}(\lambda, m) \cdot h^{(m)}_{\mathrm{FBE}}(k) + \gamma_{\mathrm{HP}} \cdot h_{\mathrm{HP}}(k) \tag{3.18}$$

for $m \in \{0, \ldots, N_{\mathrm{SE}} - 1\}$, $l \in \{1, 2\}$, and $\gamma_{\mathrm{HP}} = 0.1$. These two filters, implemented in their *non transposed form* [Oppenheim & Schafer 1995], are applied to the signal $\hat{v}_{\mathrm{hb}}(k)$ in order to obtain the shaped signal $\hat{s}_{\mathrm{hb}}(k)$. For the first 10 ms of the current frame, i.e., for $k \in \{0, \ldots, 79\}$ and $l = 1$, and, respectively, for the second half, i.e., $k \in \{80, \ldots, 159\}$ and $l = 2$, this gives

$$\hat{s}_{\mathrm{hb}}(\lambda L + k) = \sum_{j=0}^{L_{\mathrm{FBE}}-1} \hat{v}_{\mathrm{hb}}(\lambda L + k - j) \cdot h_{\mathrm{FBE},l}(\lambda, j). \tag{3.19}$$

As already mentioned in Section 2.4.4, filtering operations like (3.19) may degrade the signal's *temporal* envelope. The temporal energy distribution is potentially "smeared" over an interval which corresponds to the length of the frequency envelope shaping filter (i.e., 33 taps or 4.125 ms in this case). However, the TDBWE filterbank design ensures that this *time spread* is constrained and the signal's temporal envelope is virtually preserved. Measurements prove that for about 95% of all frames more than 90% of the energy of the impulse responses $h_{\mathrm{FBE},l}(\lambda, k)$ is concentrated within an interval of 1.375 ms. This length roughly corresponds to the temporal envelope's resolution. For the remainder of the frames, *at least* 70% of the impulse responses' energy is concentrated within this interval. Viewed from a spectral perspective, the relatively wide and overlapping frequency responses of the filterbank channels—shown in Figure 3.8—guarantee the preservation of the temporal envelope. The actual speech quality gain that is obtained with the implemented time envelope shaping can be objectively quantified, see Section 3.2.5.

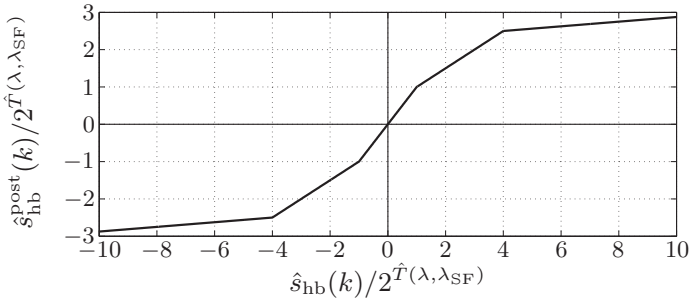


Figure 3.9: Adaptive amplitude compression function.

Adaptive Amplitude Compression

Apparently, there is no strict coupling between the TDBWE excitation $\hat{u}_{\mathrm{hb}}(k)$ and the parametric TDBWE signal description ($\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ and $\hat{F}(\lambda, m)$). Therefore, some residual artifacts may be present in the synthesized signal $\hat{s}_{\mathrm{hb}}(k)$, as, e.g., investigated in Section 2.3.2. In a CELP codec, such situations are handled by the explicit encoding of the LP residual. However, this is not possible here. Hence, to attenuate such artifacts, an *adaptive amplitude compression* is applied to $\hat{s}_{\mathrm{hb}}(k)$. Each sample within the λ_{SF}-th 1.25 ms segment of the λ-th frame is compared to the decoded and suitably aligned time envelope $\sigma \doteq 2^{\hat{T}(\lambda,\lambda_{\mathrm{SF}})}$ and the signal amplitude is compressed in order to attenuate large (short term) deviations from this envelope. As a side effect, the residual temporal smearing that is caused by the frequency envelope shaping filters can be selectively compensated. The concrete signal compression function is specified as follows:

$$\hat{s}_{\mathrm{hb}}^{\mathrm{post}}(k) = \begin{cases} \frac{\hat{s}_{\mathrm{hb}}(k)}{16} - \frac{9}{4}\sigma, & \hat{s}_{\mathrm{hb}}(k) < -4\sigma \\ \frac{\hat{s}_{\mathrm{hb}}(k)}{2} - \frac{1}{2}\sigma, & -4\sigma \leq \hat{s}_{\mathrm{hb}}(k) < -\sigma \\ \hat{s}_{\mathrm{hb}}(k), & -\sigma \leq \hat{s}_{\mathrm{hb}}(k) \leq \sigma \\ \frac{\hat{s}_{\mathrm{hb}}(k)}{2} + \frac{1}{2}\sigma, & \sigma < \hat{s}_{\mathrm{hb}}(k) \leq 4\sigma \\ \frac{\hat{s}_{\mathrm{hb}}(k)}{16} + \frac{9}{4}\sigma, & \hat{s}_{\mathrm{hb}}(k) > 4\sigma. \end{cases} \tag{3.20}$$

It is depicted in Figure 3.9.

3.2.4 Integration in the ITU-T Rec. G.729.1 Codec

The TDBWE algorithm has been standardized as a part of the ITU-T G.729.1 VoIP codec. As the TDBWE algorithm evaluation (in Section 3.2.5) has been conducted within the framework of this codec, a brief summary of the entire G.729.1 standard shall be provided here.

The G.729.1 standardization process was launched by ITU-T Study Group 16 (SG16) in May 2006 with the intention of providing a versatile and interoperable codec for wideband speech telephony and audio transmission in Voice over IP (VoIP) networks. The standardized G.729.1 speech and audio coder, a block

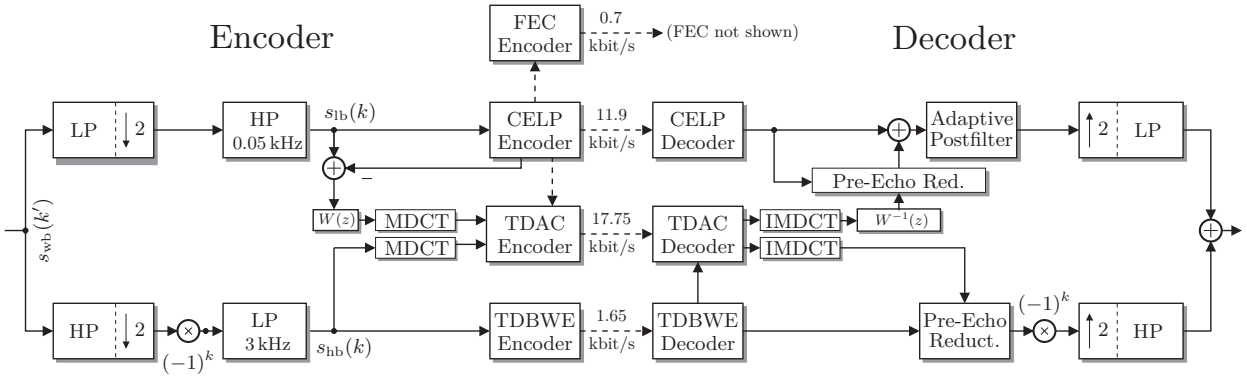


Figure 3.10: Block diagram of the ITU-T Rec. G.729.1 codec.

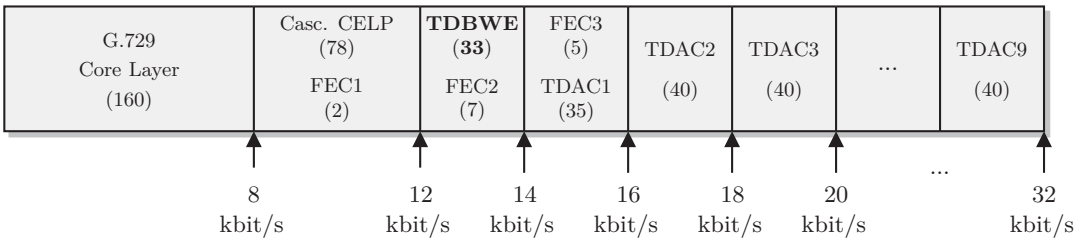


Figure 3.11: Layered bitstream format of ITU-T Rec. G.729.1. Numbers in parentheses denote bits per 20 ms frame.

diagram of which is shown in Figure 3.10, provides embedded coding with 12 bit rates between 8 and 32 kbit/s. The corresponding (layered) bitstream schematics are illustrated in Figure 3.11.

The key features of G.729.1 are, on the one hand, its interoperability with G.729 [ITU-T 1996*b*, Salami et al. 1998] which serves as the narrowband *core codec* in the embedded coding framework, and, on the other hand, its hierarchical bitstream with in total twelve bitstream layers, cf. Figure 3.11. The core codec, G.729 (with its annexes A and B), is one of the most widely deployed codecs in today's VoIP infrastructure and equipment and thus ensures interoperability as well as a basic conversational quality. Improved narrowband speech quality, especially for interconnections with traditional fixed communication networks such as ISDN (Integrated Services Digital Network) and POTS (Plain Old Telephony System) is offered with a bit rate of 12 kbit/s [Massaloux et al. 2007]. With the availability of the third bitstream layer, i.e., a total bit rate of 14 kbit/s, a wideband signal (50 Hz – 7 kHz) can be synthesized using the TDBWE technique as described above. For codec modes above 14 kbit/s, the wideband signal is further refined using an MDCT transform domain algorithm (TDAC, see also [Kövesi et al. 2004]). Additionally, a total rate of 750 bit/s is distributed over several bitstream layers to help the decoder with the concealment of lost frames, see [Vaillancourt et al. 2007].

More information on the G.729.1 codec is provided in [ITU-T 2006, Ragot et al. 2007, Varga et al. 2009] and also in [Geiser, Ragot & Taddei 2008].

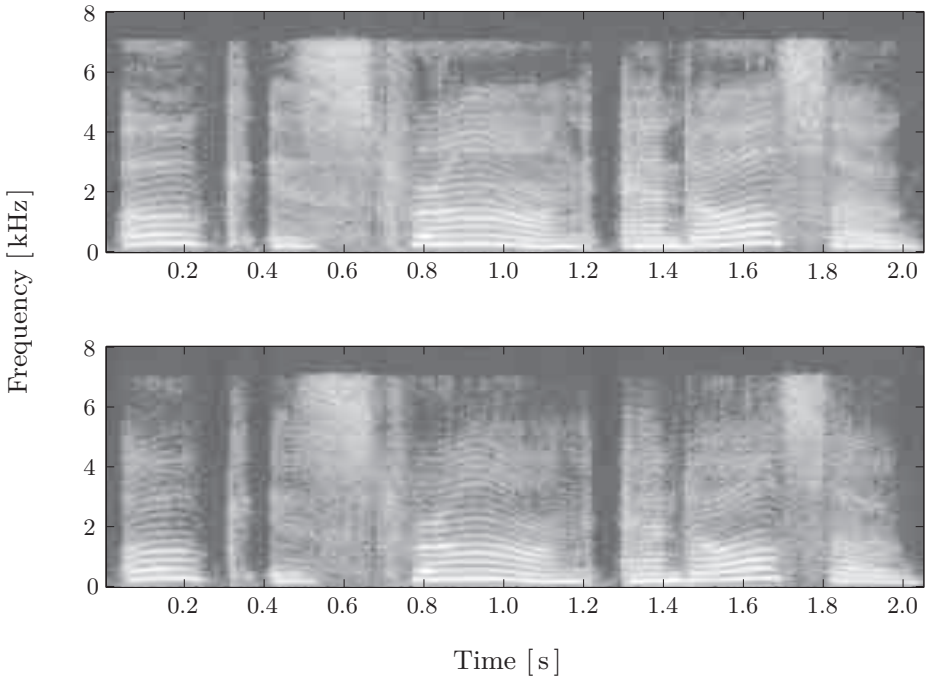


Figure 3.12: Example spectrograms of the wideband input signal (top) and of the transcoded signal (bottom, G.729.1@14 kbit/s).

3.2.5 Evaluation

The G.729.1 codec has been thoroughly evaluated and characterized in the course of the ITU-T standardization process. Here, subjective listening test results for the 14 kbit/s mode of G.729.1 are presented since this codec mode is relevant for the TDBWE performance. Additional wideband PESQ [ITU-T 2005] speech quality measurements complement the quality evaluation. The algorithmic complexity of the TDBWE implementation has been measured according to [ITU-T 1993*a*].

All tests and measurements have been conducted with the official ITU-T G.729.1 software package, i.e., a C implementation using fixed point arithmetic.

Example

As a first example, Figure 3.12 depicts two spectrograms which represent a short utterance of a female American speaker. The first spectrogram shows the original wideband input signal $s(k')$, whereas the second one is the G.729.1 output $\hat{s}(k')$ that has been decoded at a received bit rate of 14 kbit/s. In the synthetically generated high band $\hat{s}_{\mathrm{hb}}^{\mathrm{post}}(k)$, the consistent pitch structure and the properly regenerated energy envelopes are clearly visible.

Subjective Test Results

An extensive subjective quality assessment has been carried out within the "optimization and characterization phase" of the G.729.1 standardization process. An

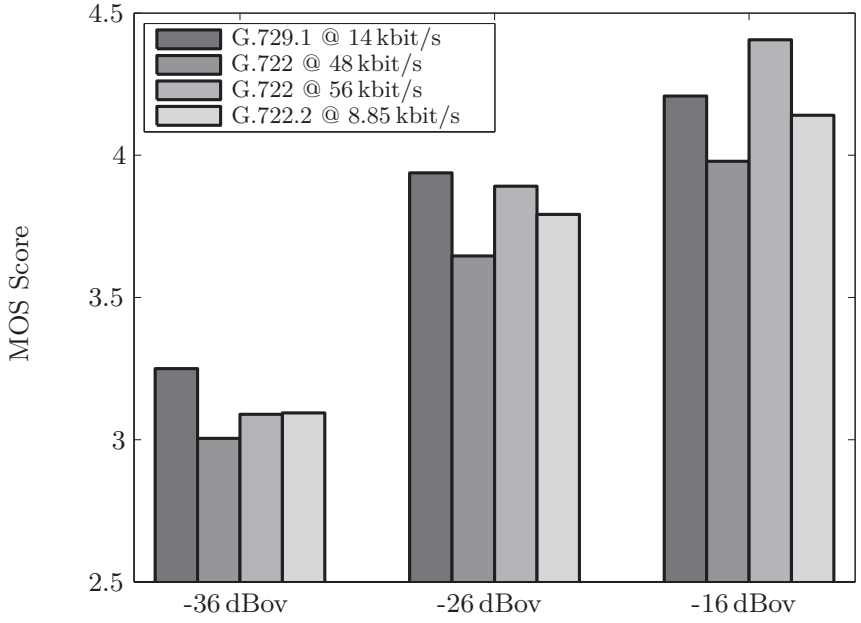


Figure 3.13: MOS scores for G.729.1 at 14 kbit/s for varying input level.

excerpt of the respective listening test results is reproduced in Figure 3.13. Note that the listening tests not only assess the quality of the high band signal, but of the entire wideband output signal, i.e., the effect of CELP encoding of the low band signal $s_{\mathrm{nb}}(k)$ is included in the rating.

In the experiment, the 14 kbit/s mode of the G.729.1 codec has been compared with other well-known references which are part of the official requirements from the ITU-T "Terms of Reference." These test references are: ITU-T Rec. G.722 at bit rates of 48 kbit/s and 56 kbit/s [ITU-T 1984] and ITU-T Rec. G.722.2 at a bit rate of 8.85 kbit/s[2] [ITU-T 2002]. As part of the test, the influence of a varying input level has been examined. The tested speech input levels are −36 dBov, −26 dBov, and −16 dBov. Thereby, "dBov" is the decibel measure w.r.t. the overload point as obtained with the ITU-T P.56 speech voltmeter, cf. [ITU-T 1993*b*, ITU-T 1993*a*].

The test items for Figure 3.13 comprised clean wideband speech signals in the English language. The listening test has been conducted using the ACR test methodology [ITU-T 1996*c*] where the 32 naïve listeners have been split into four groups of eight persons. Samples from six talkers (three male and three female) with four samples per talker (plus one sample for practice) have been presented via supra-aural headphones (closed back, e.g., Sennheiser HD25) with one capsule turned away for mono-aural listening.

An additional listening test—the results are presented in Figure 3.14—has been conducted by *France Télécom*. Here, the test objective has been to compare the

[2]ITU-T G.722.2 is identical with 3GPP AMR-WB [ETSI 2001*b*, Bessette et al. 2002].

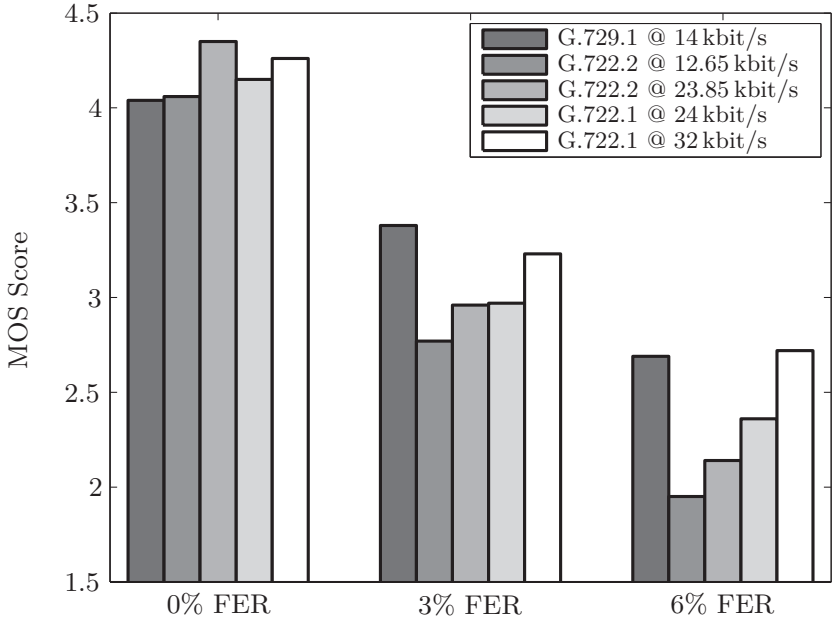


Figure 3.14: MOS scores for G.729.1 at 14 kbit/s under varying FER.

performance of G.729.1 at a bit rate of 14 kbit/s with further relevant references. Moreover, the influence of *frame erasures* was assessed. A good performance under frame erasures is crucial for the coder's targeted application in VoIP networks. In the test, the frame erasure rate (FER) has been varied between 0%, 3%, and 6%. The test laboratory used mono-aural equipment and 24 naïve listeners participated. The test samples were in French language and comprised four talkers, where four samples per talker were presented. The references for this test are ITU-T Rec. G.722.2 at bit rates of 12.65 kbit/s and 23.85 kbit/s [ITU-T 2002] and ITU-T Rec. G.722.1 at bit rates of 24 kbit/s and 32 kbit/s [ITU-T 1999].

As the main result, it could be shown that the 14 kbit/s mode of G.729.1 (including the TDBWE algorithm) is better than the G.722.2 codec at 8.85 kbit/s and almost as good as G.722.2 at 12.65 kbit/s. The first test (Figure 3.13) confirms that the quality of the G.729.1 codec is relatively stable w.r.t. varying input level. The comparatively good performance under frame erasures as shown in Figure 3.13 can mainly be attributed to the 450 bit/s of additional FEC information in Layers 2 and 3 of the G.729.1 bitstream, cf. Figure 3.11.

More listening test results for the TDBWE algorithm are presented in Chapter 6 where the concept of embedded coding is compared with parameter estimation (Chapter 4) and hidden transmission techniques (Chapter 5).

Table 3.2: Wideband PESQ measurements for G.729.1 at 14 kbit/s.

Description	Average WB-PESQ score	Standard deviation
G.729.1 14 kbit/s	3.61	0.32
without time envelope shaping	3.47	0.32
without post-processing	3.59	0.32
w/o time env. sh. & w/o post-proc.	3.40	0.31
unquantized parameter set	3.63	0.31
original high band	3.64	0.31

Objective Test Results

To analyze the performance of the TDBWE algorithm in more detail, objective quality measurements have been carried out with the wideband PESQ [ITU-T 2005] tool. The averaged wideband PESQ scores listed in Table 3.2 have been obtained from all American and British English utterances of the NTT corpus [NTT 1994].

These measurements quantify the quality gain which is obtained through the temporal envelope shaping and through the adaptive post-processing procedure (amplitude compression). Therefore, a modified codec version has been examined which skips either the temporal envelope shaping, the post-processing, or both modules. The respective wideband PESQ scores indicate that a high quality bandwidth extension can not solely rely on a spectral envelope but should also account for certain temporal signal characteristics.

Further, the validity of the TDBWE parameter quantization scheme is shown by comparing the G.729.1 wideband PESQ score with a codec version which uses the *unquantized* TDBWE parameters $T(\lambda, \lambda_{\mathrm{SF}})$ and $F(\lambda, m)$ at the decoder side. Finally, the quality for the case of a transparent (i.e., original) high band signal $s_{\mathrm{hb}}(k)$ is evaluated, whereby the low band signal $\hat{s}_{\mathrm{nb}}(k)$ is the output of G.729.1 at 12 kbit/s.

Algorithmic Complexity

Table 3.3 lists the relevant complexity figures for the TDBWE algorithm. The algorithmic complexity is—according to [ITU-T 1993*a*]—measured in weighted million operations per second (WMOPS) for the *worst case* that was observed. The complexity figures for the TDBWE part of the codec are relatively low. For the *encoder*, the major contributions come from the frequency envelope computation and from the vector quantization of the TDBWE parameters. The *decoder* complexity is dominated by the modules for excitation generation and spectral envelope shaping, respectively. Additionally, it can be observed that the TDBWE complexity

Table 3.3: Algorithmic complexity of the TDBWE algorithm.

	Module	Complexity [WMOPS]
	time envelope computation	0.03
	frequency envelope computation	0.38
	parameter quantization	0.35
	buffer handling	0.02
Σ	TDBWE encoder	0.78
	parameter decoding	0.01
	excitation generation	0.94
	time envelope shaping	0.12
	frequency envelope shaping	1.29
	adaptive amplitude compression	0.17
	buffer handling	0.01
Σ	TDBWE decoder	2.54

is asymmetrically allocated to encoder and decoder. In contrast to established speech coding algorithms like CELP, the TDBWE decoder part is considerably more complex than the encoder part.

The total TDBWE complexity amounts to 3.32 WMOPS. However, for an actual implementation of the algorithm on top of a narrowband codec, at least the band-split and the pre-processing filters (see Figures 3.2 and 3.4) have to be considered in addition to the TDBWE complexity.

Algorithmic Delay

The algorithmic delay of the entire G.729.1 codec is 48.9375 ms with contributions from framing (20 ms), QMF band-split (3.9375 ms), G.729 look-ahead (5 ms), and MDCT-window look-ahead (20 ms). In other words, the TDBWE module does *not* introduce any additional delay. The decoder-side FIR filter delay and, correspondingly, the encoder-side look-ahead of 2 ms in the frequency envelope computation are more than compensated for by the G.729 look-ahead (5 ms) in the low band branch of the codec.

Besides its "normal" mode of operation, G.729.1 offers the possibility of "low-delay" operation for its 14 kbit/s wideband mode, see Amd. 3 of the G.729.1 standard. In this mode, all MDCT domain processing in the TDAC part of the decoder is omitted and thus the algorithmic delay is reduced by the amount of the MDCT window's look-ahead, i.e., from 48.9375 ms to 28.9375 ms. Additionally, the algorithmic complexity is reduced by about 2 WMOPS.

3.2.6 Discussion

Despite its conceptual simplicity, the TDBWE algorithm proved to be a robust and flexible solution for wideband extension of narrowband speech signals. The obtained speech quality is in fact comparable to that of full-fledged wideband speech codecs. The rather low computational complexity figures make the algorithm very suitable for an implementation in portable devices. Therefore, ITU-T decided to standardize the TDBWE proposal as a part of the G.729.1 VoIP wideband codec.

Summarizing, there are several characteristics that distinguish the TDBWE algorithm from other speech bandwidth extension methods:

- The algorithm does not transmit ready-to-be-used gain factors and filter coefficients but only *desired* temporal and spectral envelopes. Gain factors and filter coefficients are *computed at the receiver*, hence the actual envelopes of the excitation signal are taken into account. This fact makes the TDBWE algorithm robust against potential deviations in the excitation signal which may, for instance, occur during and after frame losses.

- The separated analysis, transmission, and shaping of temporal and spectral envelopes make it possible to achieve a good resolution in both time and frequency domain. This leads to a good reproduction of both stationary sounds as well as transient signals. For speech signals, especially the reproduction of stop consonants and plosives benefits from the improved temporal resolution.

- The TDBWE scheme is also a very modular and flexible concept as single blocks in the receiver can easily be exchanged and improved without need to alter the encoder side or the bitstream format. Different decoders can be supported which reconstruct the wideband signal with different precision, depending on the available computational power.

- The temporal and spectral envelope parameters can not only be used for bandwidth extension purposes. In fact they may also support *subsequent signal enhancement* schemes (e.g., post-filtering and pre-/post-echo reduction [Kövesi et al. 2008]). Moreover, *additional coding stages* in a hierarchical framework, such as transform or wavelet coders, can exploit certain synergies. This has been demonstrated in [Geiser et al. 2006] and [De Meuleneire et al. 2006].

The proposed TDBWE algorithm is, owing to its speech-centric design, i.e., *speech*-trained codebooks and the assumption of certain *speech* characteristics (e.g., a unique pitch period for excitation generation), not suitable for music stimuli. To handle such situations, a much more flexible parametrization of the extension band signal must be employed. The super-wideband bandwidth extension algorithm to be described in the following section offers such flexibility, targeting generic audio signals such as music.

3.3 MDCT Domain Super-Wideband Extension

The second new bandwidth extension algorithm that is proposed in this thesis operates mainly in the frequency domain. It is designed to extend wideband audio signals (including speech and music) towards the super-wideband bandwidth.

The devised algorithm has been submitted to ITU-T by Huawei (China) and ETRI (South Korea) as a part of "Candidate Codec B" for the standardization of a new super-wideband extension of ITU-T G.729.1 [ITU-T 2006, Ragot et al. 2007] and ITU-T G.718 [ITU-T 2008*a*, Vaillancourt et al. 2008]. Various aspects of this super-wideband codec, which was in fact the only candidate to pass all requirements for mono input signals as defined in the official "Terms of Reference" (ToR), are published in [Geiser et al. 2009, Löllmann et al. 2009, Geiser, Krüger & Vary 2010, Krüger et al. 2011*b*]. The *finalized* standards were published in 2010, cf. [Laaksonen et al. 2010]. In this thesis, the focus is on the mono parts[3] of the candidate codec. The proposed bandwidth extension techniques actually play an essential role in the codec design.

In the following, first, the parameter set for the proposed super-wideband bandwidth extension algorithm is defined (Section 3.3.1). The quantization of these parameters is summarized in Section 3.3.2. The target bit rate for parameter quantization is 4 kbit/s on top of the respective wideband core codec. To properly regenerate the extension band signal in the MDCT domain, several new algorithmic measures are required in addition to the *basic* synthesis schemes that have been introduced in Chapter 2. The respective details of the MDCT domain signal synthesis are described in Section 3.3.3. The description of an integrated, novel method for joint temporal envelope control and frame erasure concealment (FEC) completes the technical details (Section 3.3.4). Afterwards, a high-level overview of the submitted "Candidate Codec B" is provided (Section 3.3.5). The respective implementation uses the G.729.1 VoIP codec as the wideband core codec. In Section 3.3.6, the proposed bandwidth extension algorithm is evaluated within the framework of this candidate codec. A discussion is finally provided in Section 3.3.7.

3.3.1 Parameter Set for Bandwidth Extension

The encoder for the 7 – 14 kHz extension band signal $s_{\mathrm{uhb}}(k)$ is illustrated in Figure 3.15. In the following, only the components which are responsible for *parametric* encoding are discussed. In the block diagram, these components are shown with solid lines. The components for *bit allocation*, *spectral normalization*, and *vector quantization* (GLCVQ) are disregarded here since they are associated with higher codec layers, i.e., higher bit rate modes (see Section 3.3.5). Moreover, only the processing for the 8 – 14 kHz extension band signal $s_{\mathrm{uhb}}(k)$ is described. Details concerning the encoding of the 7 – 8 kHz range are provided in [Geiser et al. 2009].

[3] The stereo coding aspects are beyond the scope of this thesis and, eventually, only monophonic encoding modes have been standardized by ITU-T, i.e., stereo coding has been abandoned later in the standardization process.

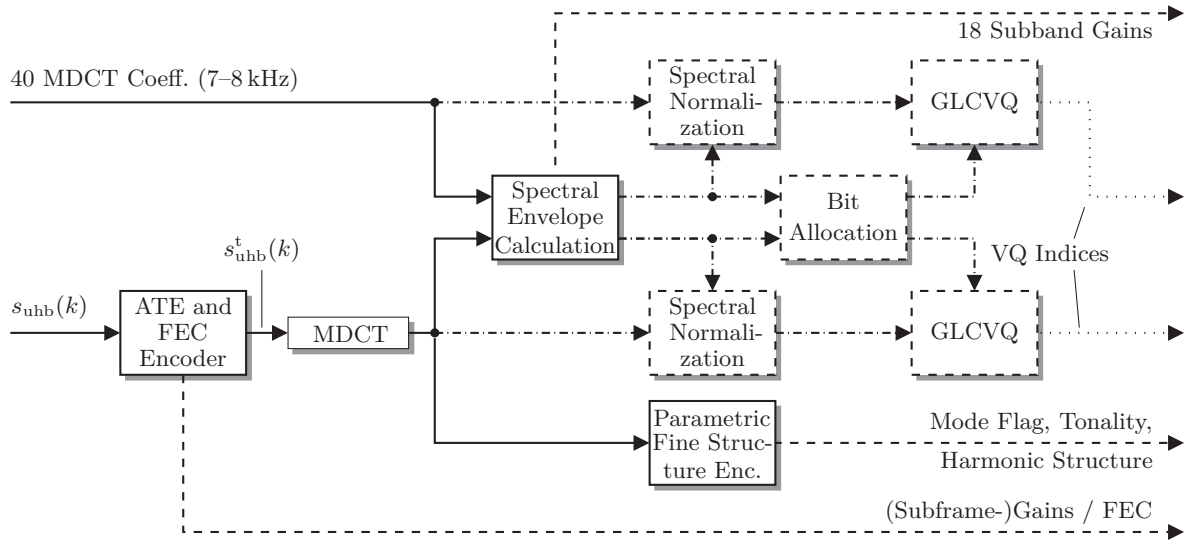


Figure 3.15: Extension Band Encoder for Super-Wideband BWE.

The *parametric* description of the extension band signal $s_{\mathrm{uhb}}(k)$ consists of an adaptive temporal envelope (ATE, cf. Section 2.2.2), a spectral subband envelope (cf. Section 2.4.2) as well as an explicit, compact description of the spectral details (cf. Section 2.5). Here, in contrast to the TDBWE algorithm from Section 3.2, the *serial* analysis approach of Figure 2.3(a) is applied, i.e., the temporal envelope is directly obtained from the time domain signal $s_{\mathrm{uhb}}(k)$ while the remaining parameters are extracted from the MDCT domain representation $S^{\mathrm{MDCT}}_{\mathrm{uhb}}(\lambda, \mu)$ of a *temporally normalized* signal $s^{\mathrm{t}}_{\mathrm{uhb}}(k)$. The computation of the individual parameters is detailed in the following.

Adaptive Temporal Envelope

The temporal envelope which is used in the present algorithm has a signal-adaptive resolution (stationary/transient frames) as introduced in Section 2.2.2, see Figure 2.5 for an example. Each 20 ms frame ($L = 320$) of the signal $s_{\mathrm{uhb}}(k)$ is divided into $N_{\mathrm{TE}} = 8$ subframes of length 2.5 ms ($L_{\mathrm{SF}} = 40$). Then, as in (3.1) for the TDBWE algorithm, logarithmic *subframe* gains with $\lambda_{\mathrm{SF}} \in \{0, \ldots, N_{\mathrm{TE}} - 1\}$ are derived:

$$T(\lambda, \lambda_{\mathrm{SF}}) = \frac{1}{2}\mathrm{ld}\,\frac{1}{L_{\mathrm{SF}}} \sum_{k=0}^{L_{\mathrm{SF}}-1} s^2_{\mathrm{uhb}}(\lambda L + \lambda_{\mathrm{SF}} L_{\mathrm{SF}} + k). \tag{3.21}$$

Additionally, a logarithmic gain is computed for the *entire* frame with index λ:

$$T(\lambda) = \frac{1}{2}\mathrm{ld}\,\frac{1}{L} \sum_{k=0}^{L-1} s^2_{\mathrm{uhb}}(\lambda L + k). \tag{3.22}$$

Based on these parameters, the temporal structure of the input signal frame is analyzed and it is classified as either "transient" ($t(\lambda) = 1$) or "stationary" ($t(\lambda) = 0$).

Therefore, a relatively simple yet effective transient detector is applied which determines if the maximum rising and/or falling slopes within the subframe gains $T(\lambda, \lambda_{\mathrm{SF}})$ exceed certain pre-specified thresholds. These thresholds have been found experimentally.

Now in *stationary* frames, only the frame gain $T(\lambda)$ is included in the parameter set for bandwidth extension whereas in *transient* frames, the N_{TE} logarithmic subframe gains $T(\lambda, \lambda_{\mathrm{SF}})$ are encoded. These parameters are quantized as described in Section 3.3.2 and used to restore the temporal signal characteristics as discussed in Section 3.3.3.

As the serial signal analysis approach for bandwidth extension (Figure 2.3(a)) is pursued here, a *temporal normalization* procedure has to be carried out before the frequency transform can be applied and the remaining parameters can be determined. Therefore, based on the quantized temporal (subframe) gains, the encoder constructs an interpolated temporal gain function (TGF) whereby the length of the interpolation window $w_{\mathrm{T}}(k)$ is adaptive. In transient frames, the TGF is constructed by an overlap-add of scaled Hann windows as shown in (2.6). During stationary frames, the employed window function is a linearly falling slope over $L/2$ samples and zero otherwise. The final gain function is then used to normalize the signal $s_{\mathrm{uhb}}(k)$

$$s^{\mathrm{t}}_{\mathrm{uhb}}(k) = s_{\mathrm{uhb}}(k) \cdot \hat{g}^{-1}_{\mathrm{TGF}}(k) \tag{3.23}$$

as in (2.7). Due to the interpolation, the TGF exhibits a pronounced lowpass characteristic such that spectral leakage is largely avoided, cf. Figure 2.4(c).

Spectral Envelope

The temporally normalized signal $s^{\mathrm{t}}_{\mathrm{uhb}}(k)$ is transformed to the frequency domain using a modified discrete cosine transform (MDCT) according to (2.20) with a Kaiser-Bessel-derived window $w_{\mathrm{F}}(k)$ of length 40 ms ($L_{\mathrm{w}} = 640$) with shape parameter $\alpha = 5$, see e.g., [Fielder et al. 1996]. For the given sampling rate of $f_{\mathrm{s}} = 16\,\mathrm{kHz}$, the transform yields 320 real-valued spectral coefficients for each 20 ms frame. Based on the 240 MDCT coefficients that correspond to the 8 – 14 kHz band of the *original* super-wideband signal, a *spectral envelope* in terms of $N_{\mathrm{SE}} = 15$ logarithmic subband gains is computed, see also (2.21). Concretely, for the m-th subband ($m \in \{0, \ldots, N_{\mathrm{SE}} - 1\}$):

$$F(\lambda, m) = \frac{1}{2}\mathrm{ld}\,\frac{1}{M_{\mathrm{SB}}} \sum_{\mu=0}^{M_{\mathrm{SB}}-1} W(\mu) \cdot |S_{\mathrm{uhb}}(\lambda, \mu + mM_{\mathrm{s}})|^2 \tag{3.24}$$

with $M_{\mathrm{SB}} = M_{\mathrm{s}} = 16$. The number of subbands (15) might appear unnecessarily high when compared with the results of Section 2.6. However, in the super-wideband codec, the gains $F(\lambda, m)$ shall be reused as scale factors for spherical vector quantization (e.g., [Krüger et al. 2008]) which requires a higher number of subbands. A *rectangular* window function $W(\mu)$ is used here for the same reason.

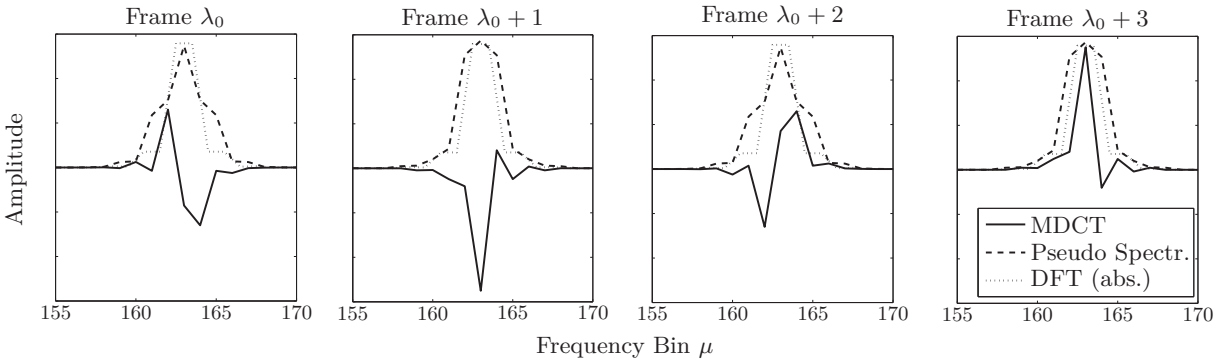


Figure 3.16: MDCT, DFT, and pseudospectrum of a stationary sinusoid over four consecutive analysis frames.

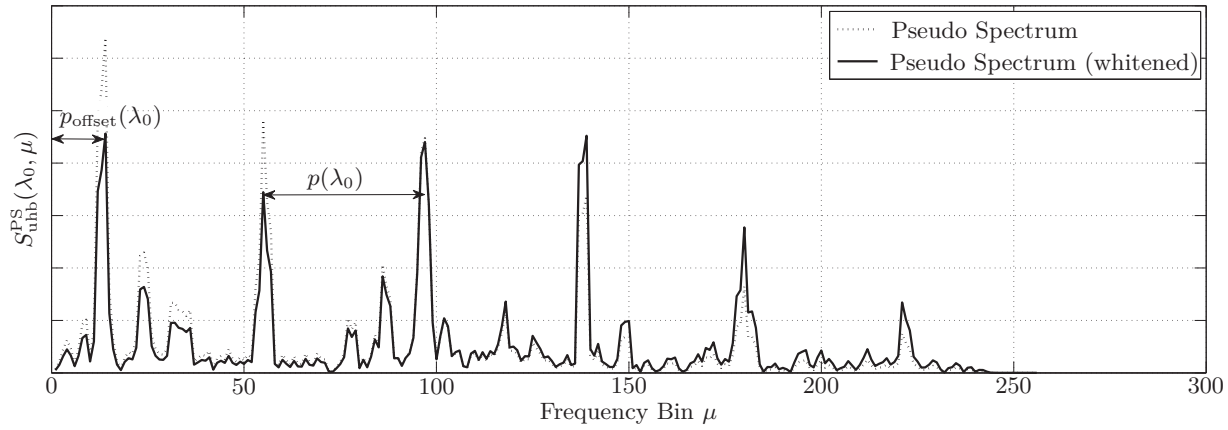


Figure 3.17: Pseudo spectrum of a harmonic signal in the 8 – 14 kHz extension band.

Parametric Encoding of Spectral Details

As opposed to the TDBWE algorithm, the present super-wideband extension explicitly encodes the *spectral details* of the extension band signal with a few bits per frame to supply a coarse description/classification thereof to the decoder. The computation of these parameters is briefly outlined in the following.

For the parameter set to be derived, the *harmonic structure* of the 8 – 14 kHz MDCT signal is analyzed. However, it is important to note that the MDCT as such is not well suited for spectral analysis. This fact is easily illustrated with the MDCT representation of a stationary sinusoid signal as shown in Figure 3.16 which is, obviously, *not* stationary. A concise spectral analysis is therefore difficult to achieve. To avoid an additional (and costly) DFT for spectral analysis, it is proposed here to use the so called "pseudo spectrum" (PS) representation [Daudet & Sandler 2004] instead:

$$S^{\text{PS}}_{\text{uhb}}(\lambda, \mu + mM_{\text{s}}) = \sqrt{S_{\text{uhb}}(\lambda, \mu - 1)^2 + (S_{\text{uhb}}(\lambda, \mu) - S_{\text{uhb}}(\lambda, \mu + 1))^2}. \quad (3.25)$$

The MDCT pseudo spectrum approximates the DFT amplitude spectrum (dotted graph in Figure 3.16) solely based on the MDCT coefficients of the *current* frame.

For practical reasons, an additional spectral whitening (tilt compensation) is applied in the codec. From the pseudo spectrum $S_{\mathrm{uhb}}^{\mathrm{PS}}(\lambda, \mu)$, the following parameters are derived:

- The dominating harmonic pitch frequency is identified by computing a *harmonic grid* parameter $p(\lambda)$, which is given in multiples of one MDCT frequency bin. The parameter is obtained by identifying the first local maximum of the (interpolated) autocorrelation function (ACF) $\varphi_{SS}(\mu)$ of the pseudo spectrum vector. Lower values for the pitch frequency are preferred to avoid pitch doubling errors. The frequency resolution of the pitch search algorithm is 6 Hz. As the employed MDCT with $L_{\mathrm{w}}/2 = 320$ only provides a spectral resolution of 25 Hz, an oversampling factor of four has to be used.

- The direct extrapolation of *higher order* harmonics in the extension band is difficult because any inaccuracy in the harmonic grid estimation is multiplied. Therefore, in addition to the grid parameter, a *harmonic offset* parameter $p_{\mathrm{offset}}(\lambda)$ is needed. The pitch harmonics in the (downsampled) 8–14 kHz subband are equally spaced, but they are, in general, *not* placed at an integer multiple of the fundamental frequency, see Figure 3.17 for an illustration.

- An inverse *tonality* value $\tau(\lambda) \in [0, 1]$ for the extension band is computed based on the ACF $\varphi_{SS}(\mu)$ of the pseudo spectrum $S_{\mathrm{uhb}}^{\mathrm{PS}}(\lambda, \mu)$:

$$\tau = \min_{\mu \in \{\mu_{\min}, \ldots, \mu_{\max}\}} \frac{\varphi_{SS}(\mu)}{\varphi_{SS}(0)}. \tag{3.26}$$

  The search range $\{\mu_{\min}, \ldots, \mu_{\max}\}$ for the tonality parameter is identical to the search range of the harmonic grid parameter $p(\lambda)$.

- Finally, a binary flag $f(\lambda)$ is determined to select the synthesis mode for the spectral details at the decoder, i.e., "spectral replication" or "harmonic synthesis."

  The flag is set to 1 (spectral replication) if either the pitch frequency of the baseband CELP codec is close to the identified extension band pitch frequency (inverse harmonic grid $p(\lambda)$) or if the identified extension band pitch frequency is below 400 Hz. In this case, the details of the baseband signal are considered to be "similar" to the extension band details. Therefore, the harmonic grid and offset parameters are not transmitted.

  On the contrary, $f(\lambda)$ is set to 0 (harmonic synthesis) if the identified pitch frequency is higher than 400 Hz and if the CELP codec parameters do not adequately represent the harmonic structure in the extension band. In this case, the details of the baseband signal are considered to be "dissimilar" to the extension band details. Hence, the harmonic grid and offset parameters are actually transmitted in the bitstream.

3.3.2 Quantization

In the encoder, the parameter set for super-wideband bandwidth extension, i.e., the adaptive temporal envelope, the spectral envelope and the coarse description of spectral details, is quantized and added to the bitstream in the form of an additional bitstream layer. The quantization of the individual parameters is detailed in the following sections. The consumed bit rate ranges from 2.7 kbit/s to 4.75 kbit/s depending on the characteristics of the current extension band input frame.

Adaptive Temporal Envelope (and FEC Information)

The transmission of an *adaptive* temporal envelope (ATE) requires an additional binary flag ($t(\lambda)$, 1 bit/20 ms = 50 bit/s) to indicate transient signal frames. As the ATE information shall also be reusable for purposes of frame erasure concealment (FEC) at the decoder side, a special bitstream arrangement is required. For a *signal frame* with a given index λ, the corresponding "bitstream frame" with index λ is composed as follows:

- $t(\lambda)$ — The transient flag of the current frame.
- $t(\lambda-1)$ — The (repeated) transient flag of the previous frame.
- If $t(\lambda) = 0$: $\hat{T}(\lambda)$ — The gain of the current frame, quantized with a wordlength of 5 bits in the logarithmic domain.
- If $t(\lambda-1) = 0$: $\hat{T}(\lambda-1)$ — The (repeated) gain of the previous frame, quantized with a wordlength of 5 bits in the logarithmic domain.
- If $t(\lambda) = 1$: $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{TE}}-1\}$ — The encoded subframe gains with *odd* indices corresponding to the *current* frame; the first subframe gain $\hat{T}(\lambda, 1)$ is quantized with a wordlength of 5 bits while each of the subsequent *differential* subframe gains $T(\lambda, 3) - \hat{T}(\lambda, 1)$, $T(\lambda, 5) - \hat{T}(\lambda, 3)$, and $T(\lambda, 7) - \hat{T}(\lambda, 5)$ is encoded with 4 bits.
- If $t(\lambda-1) = 1$: $\hat{T}(\lambda-1, \lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{0, 2, \ldots, N_{\mathrm{TE}}-2\}$ — The encoded subframe gains with *even* indices corresponding to the *previous* frame; the first subframe gain $\hat{T}(\lambda, 2)$ is quantized with a wordlength of 5 bits while each of the subsequent *differential* subframe gains $T(\lambda, 4) - \hat{T}(\lambda, 2)$, $T(\lambda, 6) - \hat{T}(\lambda, 4)$, and $T(\lambda, 8) - \hat{T}(\lambda, 6)$ is encoded with 4 bits.

Note that this bitstream arrangement does not introduce any additional delay into the codec, because, in any case, the overlap-add operation of the inverse MDCT requires one frame of delay *at the decoder side*, see also Section 2.2.3.

The proposed bitstream arrangement is further illustrated in Table 3.4 which shows a concrete example for a stationary/transient/stationary frame sequence. The information is either transmitted redundantly (for stationary frames, e.g.,

Table 3.4: Adaptive temporal envelope encoding: Bitstream arrangement for a stationary/transient/stationary frame sequence. The number of bits for each parameter is shown in parentheses.

$\lambda = 1$		$\lambda = 2$		$\lambda = 3$	
$t(1) = 0$	(1)	$t(2) = 1$	(1)	$t(3) = 0$	(1)
$t(0) = 0$	(1)	$t(1) = 0$	(1)	$t(2) = 1$	(1)
$\hat{T}(1)$	(5)	–		$\hat{T}(3)$	(5)
$\hat{T}(0)$	(5)	$\hat{T}(1)$	(5)	–	
–		$\hat{T}(2,1), \hat{T}(2,3), \ldots$	(17)	–	
–		–		$\hat{T}(2,0), \hat{T}(2,2), \ldots$	(17)
Σ	12 bits	Σ	24 bits	Σ	24 bits

$\lambda = 1$) or distributed across neighboring bitstream frames (for transients, $\lambda = 2$). The amount of transmitted redundancy is kept to a minimum and, in particular for transient frames, there is no redundant information (except for the repeated mode bit). The total number of bits for the adaptive temporal envelope parameters per 20 ms-frame is 12 for consecutive stationary frames, 24 for transient-stationary transitions and 36 for consecutive transient frames (0.6, 1.2, 1.8 kbit/s).

The actual algorithm to conceal lost frames is detailed in Section 3.3.4, where it will become clear that the proposed bitstream arrangement of the temporal envelope parameters can be elegantly reused for frame erasure concealment.

Spectral Envelope

The quantization of the spectral envelope parameters $F(\lambda, m)$ with $m \in \{0, \ldots, 14\}$ is, as the TDBWE parameter quantization (Section 3.2.2), realized by mean-removed vector quantization. However, the speech-specific codebooks of the TDBWE algorithm are replaced by a more generic *spherical* vector codebook. Therefore, the spectral envelope of generic audio signals can be better represented and encoded with a sufficient quality.

For spherical quantization, the *Gosset Low Complexity Vector Quantizer* (GLCVQ) as described in [Krüger et al. 2011*b*] is used. The total bit rate for the vector mean, radius, and shape (direction) parameters ranges from 38 to 41 bit/frame (1.9 – 2.05 kbit/s) depending on the available bit budget.

In the codec, the spherical vector quantization procedure is followed by an optional *scalar* quantization with entropy coding of the residual quantization error, see [Geiser et al. 2009]. This encoding module is not detailed here for brevity.

Spectral Details

The parameter quantization for the spectral details of the extension band is dependent on the *characteristics* of the current input frame as detailed in Table 3.5.

Table 3.5: Bit allocation for the encoding of the spectral details. The number of bits per frame depends on the *characteristics* of the input signal, i.e., on certain conditions for $f(\lambda)$ and $\hat{\tau}(\lambda)$. The thresholds τ_1 and τ_2 are fixed.

Condition	$f(\lambda) = 1$ (spec. rep.)	$f(\lambda) = 0$ $\hat{\tau}(\lambda) > \tau_1$	$f(\lambda) = 0$ $\hat{\tau}(\lambda) < \tau_2$	$f(\lambda) = 0$ $\tau_2 \leq \hat{\tau}(\lambda) \leq \tau_1$
$\hat{\tau}(\lambda)$	3	3	3	3
$f(\lambda)$	1	1	1	1
$p(\lambda)$ (integer)	-	-	6	6
$p(\lambda)$ (frac.)	-	-	2	-
$p_{\text{offset}}(\lambda)$	-	-	ld $\lceil p(\lambda) \rceil$ (max. 6)	ld $\lceil p(\lambda) \rceil$ (max. 6)
Σ	4 bits	4 bits	max. 18 bits	max. 16 bits
Bit rate	0.2 kbit/s	0.2 kbit/s	max. 0.9 kbit/s	max. 0.8 kbit/s

The inverse tonality value $\hat{\tau}(\lambda)$ for the extension band as well as the binary mode flag $f(\lambda)$ (determining whether spectral replication or harmonic synthesis shall be applied) are *always* included in the bitstream. The transmission of the remaining parameters, i.e., of the integer and fractional part of the harmonic grid $p(\lambda)$ as well as of the harmonic offset $p_{\text{offset}}(\lambda)$, depends on the binary flag $f(\lambda)$ and on the actual value of $\hat{\tau}(\lambda)$.

In total, between 4 and 18 bits (0.2 – 0.9 kbit/s) are consumed by this description of the spectral details.[4] This way, the bit rate is adapted to the amount of harmonicity that is present in the extension band.

3.3.3 Synthesis

The synthesis for the super-wideband extension band signal $\hat{s}_{\text{uhb}}(k)$ encompasses the solid (non-dashed) blocks of Figure 3.18. These components are described in the following.

Regeneration of the Spectral Fine Structure

As the first algorithmic step in the parametric signal synthesis, the spectral fine structure of the extension band signal is regenerated directly in the MDCT domain. The concrete mode of operation depends on the transmitted binary flag $f(\lambda)$.

If $f(\lambda)$ is set to 1, the fine structure $\hat{U}_{\text{uhb}}(\lambda, \mu)$ in the 8 – 14 kHz frequency range is derived by spectral replication from the 1 – 7 kHz band of $\hat{S}_{\text{wb}}(\lambda, \mu)$, see Section 2.5.1 for a description of the replication procedure. Additionally, the

[4] It should be noted that the parameters related to the harmonic structure might not fit in the first 4 kbit/s bitstream enhancement layer. They are simply omitted in these cases.

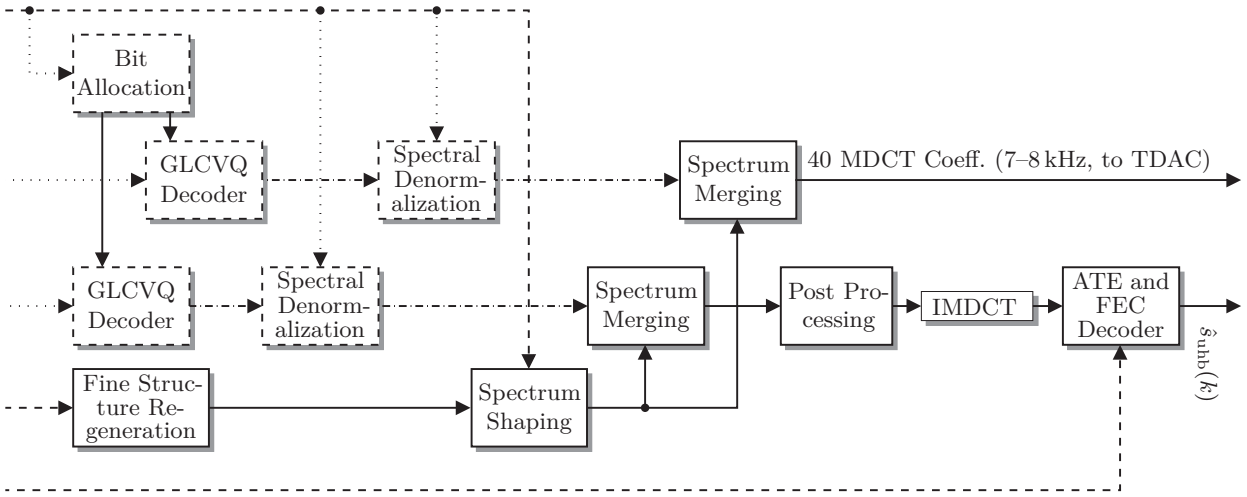


Figure 3.18: Extension Band Decoder for Super-Wideband BWE.

tonality of the replicated signal is adjusted according to the received tonality value $\hat{\tau}(\lambda)$. This is achieved by emphasizing spectral peaks or by adding pseudo random noise, respectively.

If $f(\lambda)$ is set to 0, the fine structure in the 8 – 14 kHz frequency range is generated as a mixture of pseudo random noise and synthetic sinusoidal components (harmonics). The energy ratio of noisy and of the harmonic components is controlled by the inverse tonality value $\hat{\tau}(\lambda)$.

Yet, as already indicated by the instationarity of the transform coefficients for a stationary input signal (see Figure 3.16), sinusoidal synthesis in the MDCT domain is not straightforward. For example, placing unit peaks in the MDCT domain (and keeping this set of transform coefficients *constant* over several frames) leads to a *temporally modulated* time domain signal that is affected by annoying artifacts, cf. [Daudet & Sandler 2004]. Therefore, to obtain concise sinusoids in the time domain, a new method is proposed here in which the individual harmonic components are synthesized by *imitating* the time-variant MDCT domain behavior as shown in Figure 3.16. The individual pitch harmonics are generated by placing the correct sequence of MDCT domain pulse shape prototypes at the appropriate positions in the transform domain representation (MDCT bins μ). The concrete positions are derived from the received[5] pitch grid and pitch offset parameters $\hat{p}(\lambda)$ and $\hat{p}_{\mathrm{offset}}(\lambda)$.

An example synthetic MDCT spectrum $\hat{U}_{\mathrm{uhb}}(\lambda, \mu)$ with strong harmonic components is shown in Figure 3.19(b). Compared to the original input spectrum in Figure 3.19(a), the pitch grid and the offset are correctly restored, but the phase information (corresponding to the pulse shape in the MDCT domain, cf. Figure 3.16) is not preserved. Nevertheless, at least a *continuous phase evolution* over successive signal frames λ is guaranteed.

[5] If these parameters have not been transmitted in the bitstream (see Section 3.3.2), they are extrapolated from the baseband CELP codec.

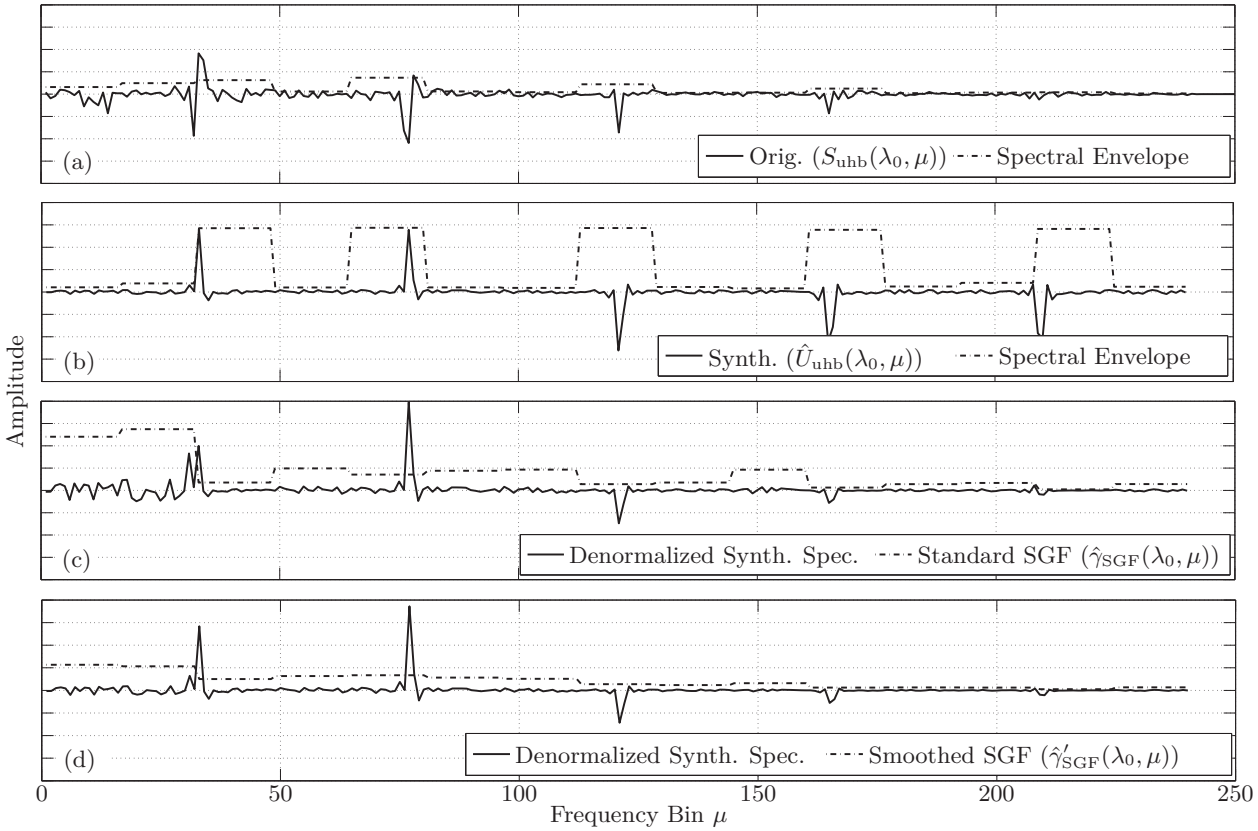


Figure 3.19: MDCT domain harmonic synthesis and spectral shaping.

Spectral Envelope

To restore the spectral envelope of the extension band signal, the regenerated MDCT coefficients $\hat{U}_{\mathrm{uhb}}(\lambda,\mu)$ are spectrally shaped by subband-wise multiplication with appropriate gain correction factors. Following Section 2.4.3, these factors are the ratio of the desired subband gains $\hat{\gamma}(\lambda,m) = 2^{\hat{F}(\lambda,m)}$ (spectral envelope) and of the measured gains $\hat{\gamma}'(\lambda,m)$ of the regenerated fine structure coefficients. Note that no interpolation window $W_{\mathrm{S}}(\mu)$ is applied here.

However, there is a specific exception that occurs with very tonal subbands: If a tonal component is located very close to a subband boundary, *interpolated gain factors* are applied, whereby the interpolation is applied over the subband index m. The intention is to mitigate potential artifacts due to slightly misplaced harmonics, in particular for high pitched signals (e.g., violin). The underlying effect is illustrated in Figure 3.19. Note that in particular the first harmonic component is located very close to a subband boundary ($\mu = 32$). Figures 3.19(c) and (d) illustrate the restoration of the spectral envelope by using the standard method and by using interpolated gain factors, respectively. In Figure 3.19(c) it can be seen that the coefficients of the *second* subband ($\mu \in \{16,\ldots,31\}$) are multiplied with an exceedingly high gain factor, leading to unwanted noise amplification. The application of interpolated gain factors reduces this effect, resulting in a much more consistent regeneration of the respective harmonic components.

MDCT Domain Post-Processing and Inverse Transform

After the restoration of the spectral envelope, MDCT-domain post-processing similar to the method of [ITU-T 2006] is applied and the resulting spectrum is transformed into the time domain according to (2.27). Then, with the overlap-add operation (2.28), the (still temporally normalized) signal $\hat{s}^{\mathrm{t}}_{\mathrm{uhb}}(k)$ is obtained.

Adaptive Temporal Envelope

The final algorithmic step to regenerate the extension band signal $\hat{s}_{\mathrm{uhb}}(k)$ is, according to Figure 3.18, the restoration of the temporal envelope contour:

$$\hat{s}_{\mathrm{uhb}}(k) = \hat{s}^{\mathrm{t}}_{\mathrm{uhb}}(k) \cdot \hat{g}_{\mathrm{TGF}}(k). \tag{3.27}$$

The gain function $\hat{g}_{\mathrm{TGF}}(k)$ for temporal denormalization is constructed from the *received* (sub)frame gains $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ or $\hat{T}(\lambda)$, respectively. The construction itself follows the description from the encoder side, i.e., an adaptive interpolation window $w_{\mathrm{T}}(k)$ is used for this purpose, see Section 3.3.1. Owing to this stationary/transient distinction, the spectral details in stationary segments can be well preserved.

The proposed temporal denormalization procedure is able to effectively suppress so-called pre-echo artifacts which are very common in transform audio coding and that are particularly strong in the case of a purely parametric signal synthesis [Geiser & Vary 2009]. To illustrate the effect, an example signal has been synthesized with and without explicit temporal envelope control, see Figure 3.20(a) – (c). The temporal envelope in Figure 3.20(c) is clearly improved in comparison to that of Figure 3.20(b).

In the following, a new method for frame erasure concealment is described. It is based on the same parameters as the temporal gain control mechanism.

3.3.4 Frame Erasure Concealment

As already outlined in Section 2.2.3, the temporal envelope control according to (3.27) can be used to realize a novel mechanism for *frame erasure concealment* (FEC). The basis for this algorithm is the special bitstream arrangement that has been described in Section 3.3.2 and in Table 3.4. It is important to note that the algorithm does not incur any additional algorithmic delay.

If a *frame erasure* is signaled to the decoder, the decoded signal $\hat{s}^{\mathrm{t}}(\lambda L + k)$ in the current frame is unavailable and has to be estimated. The simplest approach is frame repetition: $\tilde{s}^{\mathrm{t}}(\lambda L + k) = \hat{s}^{\mathrm{t}}((\lambda - 1)L + k)$. Though, a better alternative is to repeat the *transform* coefficients instead, i.e., the MDCT coefficients of the spectral details $\tilde{U}_{\mathrm{uhb}}(\lambda, \mu)$ are copied from the coefficient set $\hat{U}_{\mathrm{uhb}}(\lambda - 1, \mu)$ of the *previous* frame. The inverse transform and the overlap-add operation will then smooth the transition between the (possibly correct) previous frame and the missing frame. Now, with the estimated signal $\tilde{s}^{\mathrm{t}}(\lambda L + k)$ and with the (partially) available temporal envelope information, the concealed output can be produced.

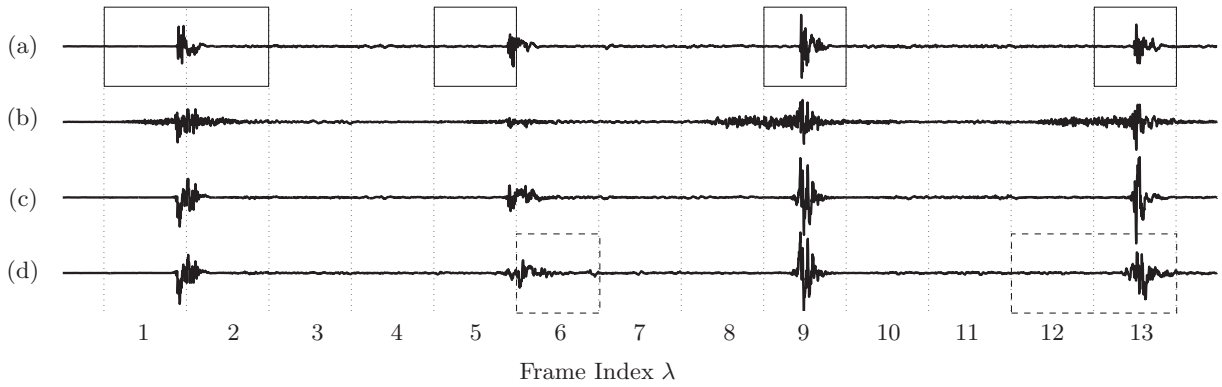


Figure 3.20: Example waveforms for the proposed method for joint temporal envelope control and frame erasure concealment.
(a) Original extension band signal (castanets),
Solid boxes: frames classified as transient ($t(\lambda) = 1$).
(b) Synthesized signal *without* temporal envelope control.
The subjectively objectionable pre- and post echo artifacts are clearly visible.
(c) Synthesized signal with temporal envelope control ($\hat{s}(k)$).
(d) Synthesized signal with concealed frame erasures ($\tilde{s}(k)$).
All bits from the frames with dashed boxes have been discarded from the bitstream.

Note that, for a *single* lost frame, the transient/stationary flag $t(\lambda)$ is still available from the *subsequent* bitstream frame (index λ+1), cf. Section 3.3.2. Depending on the identified frame type (as determined by $t(\lambda)$), the TGF $\hat{g}_{\mathrm{TGF}}(\lambda L + k)$ is now reconstructed in different ways:

- For missing *stationary* frames ($t(\lambda) = 0$), the overall gain factor $\hat{T}(\lambda)$ is available from the redundant bitstream.

  If, in the example from Table 3.4, the decoder wanted to reconstruct signal frame $\lambda = 1$ but the corresponding frame in the bitstream was lost, the flag $t(1) = 0$ and the respective gain $\hat{T}(1)$ would still be available from the second bitstream frame ($\lambda = 2$).

- When a *transient* frame is lost ($t(\lambda) = 1$), the information about the gains of the *subframes* with *odd* indices $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ (where $\lambda_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$) is missing. Therefore, they have to be interpolated from the subframe gains with *even* indices: $\tilde{T}(\lambda, \lambda_{\mathrm{SF}}) = \left[\hat{T}(\lambda, \lambda_{\mathrm{SF}} - 1) + \hat{T}(\lambda, \lambda_{\mathrm{SF}} + 1)\right]/2$ (where $\lambda_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$). Then, a TGF (with reduced temporal resolution) can be constructed.

  If, in the example from Table 3.4, the decoder wanted to reconstruct signal frame $\lambda = 2$, but the corresponding frame in the bitstream was lost, the subframe gains with even indices would still be available from the third

frame ($\lambda = 3$) and the gains with odd indices could be obtained through interpolation.

- Likewise, if a frame *after* a transient frame is lost, the *even* subframe gains (describing the *preceding* transient) are unavailable. Again, the complete TGF cannot be reconstructed and only the subframe gains with *odd* indices, i.e., $\hat{T}(\lambda, \lambda_{\mathrm{SF}})$ with $\lambda_{\mathrm{SF}} \in \{1, 3, \ldots, N_{\mathrm{SF}} - 1\}$, are used to form the TGF. Therefore, the subframe gains with *even* indices are obtained through interpolation according to $\tilde{T}(\lambda, \lambda_{\mathrm{SF}}) = \left[\hat{T}(\lambda, \lambda_{\mathrm{SF}} - 1) + \hat{T}(\lambda, \lambda_{\mathrm{SF}} + 1)\right]/2$, whereby $\lambda_{\mathrm{SF}} \in \{0, 2, \ldots, N_{\mathrm{SF}} - 2\}$ in this case.

  If, again in the example from Table 3.4, the decoder wanted to reconstruct signal frame $\lambda = 2$ and the *third* bitstream frame ($\lambda = 3$) was lost, the subframe gains with odd indices would already be available from bitstream frame $\lambda = 2$ and the missing gains with even indices could be obtained through interpolation.

- In case of *bursty frame erasures*, i.e., if two or more consecutive bitstream frames are lost, there is no information available about the current decoder frame. Therefore, the decoder has to resort to an extrapolation based approach where the currently missing frame is assumed to be stationary and the (averaged) gain factor from the previous (reconstructed) frame is decreased by a certain amount before applying the gain denormalization.

Figure 3.20 illustrates the performance of the proposed method for joint temporal envelope control and frame erasure concealment based on the 8 – 14 kHz components of the EBU SQAM castanet signal [EBU 1988]. The G.729.1-SWB codec has been run at its lowest bit rate (4 kbit/s on top of G.729.1). A considerable amount of pre-echoes is actually produced as seen in Figure 3.20(b). When applying the TGF, these artifacts are clearly reduced (Figure 3.20(c)). A signal with *concealed* frames is shown in Figure 3.20(d). In this example, the bits from Frames 6, 12 and 13 have been discarded. Note that also Frame 5 is a "partially concealed" output since the subframe gains with even indices from Frame 6 are missing.

3.3.5 Integration in the ITU-T G.729.1-SWB Candidate Codec

The proposed super-wideband bandwidth extension algorithm has been developed in the course of an ITU-T standardization project. The respective candidate codec, which was submitted to ITU-T by Huawei (China) and ETRI (South Korea), is based on the ITU-T G.729.1 VoIP wideband codec (see Section 3.2.4 for a description). Block diagrams of the encoder and decoder are shown in Figure 3.21 and Figure 3.22, respectively. Here, a brief summary of the submitted codec ("Candidate Codec B") and of its capabilities shall be provided.

The standardization project was launched in 2008 when the demand for embedded speech and audio codecs for conversational applications that can offer an even

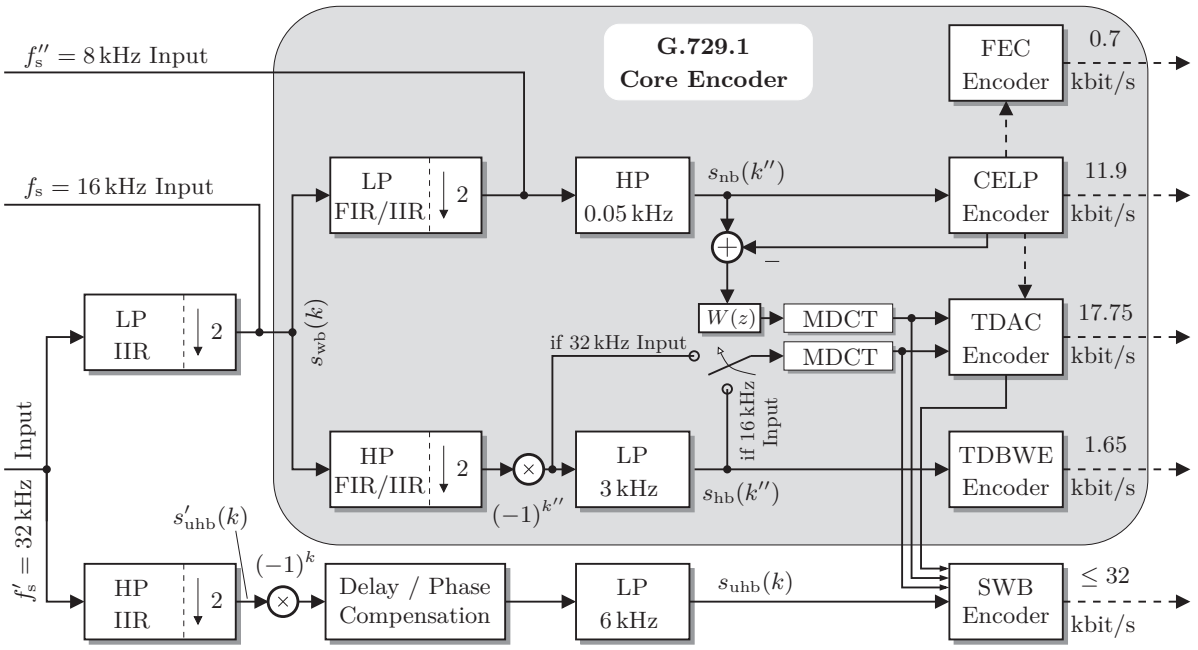


Figure 3.21: G.729.1-SWB Candidate B — Encoder

better quality than previous wideband speech codecs has been recognized within ITU-T Study Group 16 (SG16). It was required to enhance both the ITU-T G.729.1 [ITU-T 2006, Ragot et al. 2007] and ITU-T G.718 [ITU-T 2008*a*, Vaillancourt et al. 2008] wideband VoIP codecs with the ability to encode generic *audio* signals with a *super-wideband* audio bandwidth.

The proposed super-wideband extension of G.729.1 increases the transmitted frequency range from 7 kHz to 14 kHz. Therefore, the maximum bit rate is raised from 32 kbit/s to 64 kbit/s by adding five layers to the standardized G.729.1 bitstream (see Figure 3.11). In the context of the super-wideband codec, the layered G.729.1 bitstream can itself be viewed as the new "core layer" with a bit rate of 32 kbit/s. On top of that, there are five "enhancement layers," the first two add 4 kbit/s each and the latter three add 8 kbit/s each to the bitstream. Thereby, the first (and in some cases, partially, the second) bitstream enhancement layer, comprise the information which can be used for parametric bandwidth extension towards super-wideband according to Sections 3.3.1 – 3.3.4. The remaining enhancement layers are based on conventional transform coding techniques where groups of spectral coefficients are quantized with a spherical vector quantizer. The new *Gosset Low Complexity Vector Quantizer* (GLCVQ) which is described in [Krüger et al. 2010, Krüger et al. 2011*a*, Krüger et al. 2011*b*] is used for this purpose. More details concerning the proposed candidate codec can be found in [Geiser et al. 2009, Löllmann et al. 2009, Geiser, Krüger & Vary 2010, Krüger et al. 2011*b*].

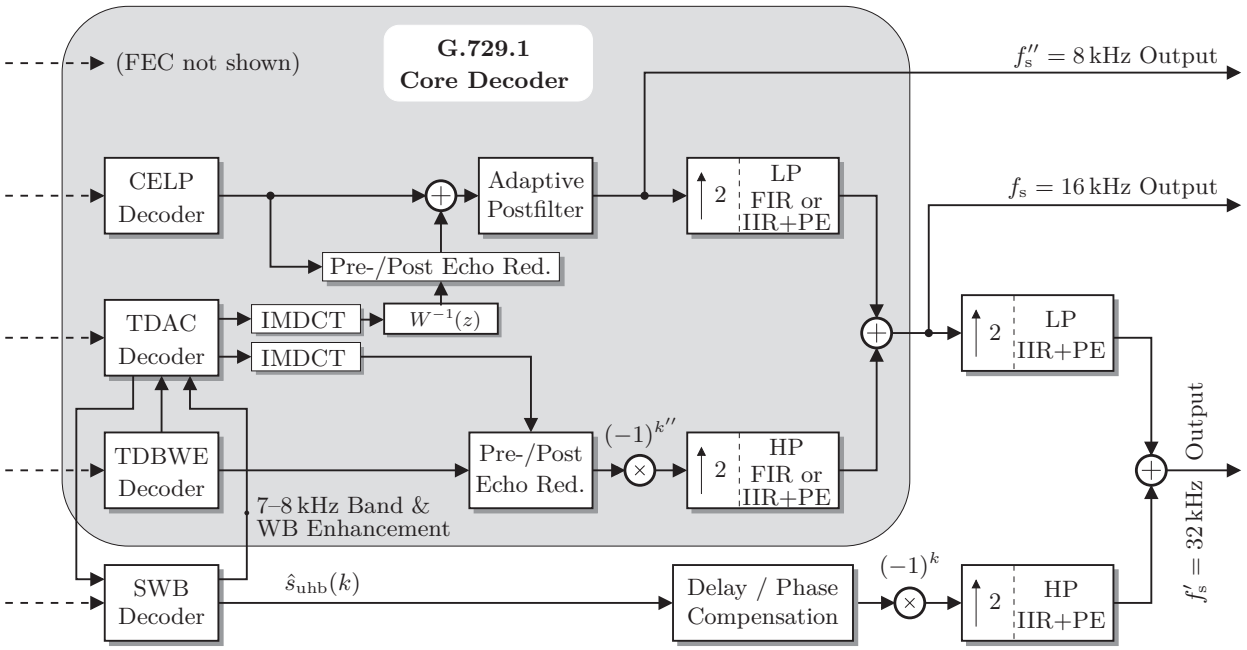


Figure 3.22: G.729.1-SWB Candidate B — Decoder

3.3.6 Evaluation

The described bandwidth extension algorithm for super-wideband speech and audio signals is evaluated in this section. It is important to recall that the algorithm is part of a complete super-wideband codec which, at higher bit rates, also provides *quantized* spectral coefficients (see Figure 3.22). The test conditions that are discussed in the following are specific coding modes of this codec, concretely the 36 kbit/s and 40 kbit/s modes are used. These modes add 4 or 8 kbit/s on top of the G.729.1 wideband bitstream.

In the 40 kbit/s mode, the added information includes the previously discussed bandwidth extension parameters and, if possible, a refinement of the spectral envelope parameters which is based on scalar quantization and Huffman coding. With the additional 8 kbit/s, in some cases, the bit budget already allows for the transmission of a few *quantized* spectral coefficients. In such cases, the coding scheme is, strictly speaking, not purely parametric anymore, although the actual influence of only a few quantized coefficients on the obtained quality is very limited. The 36 kbit/s mode, in contrast, is guaranteed to deliver a purely parametric resynthesis of the extension band signal. Here, however, a few bandwidth extension parameters (harmonic pitch and grid) may have been discarded because of the limited bit budget.

All presented evaluation results have been obtained with the software package that has been submitted to ITU-T for qualification as "Candidate Codec B." If possible, the official qualification test results are reproduced. The softwarea implementation in C has been instrumented with additional complexity counters. It is based on floating point arithmetic.

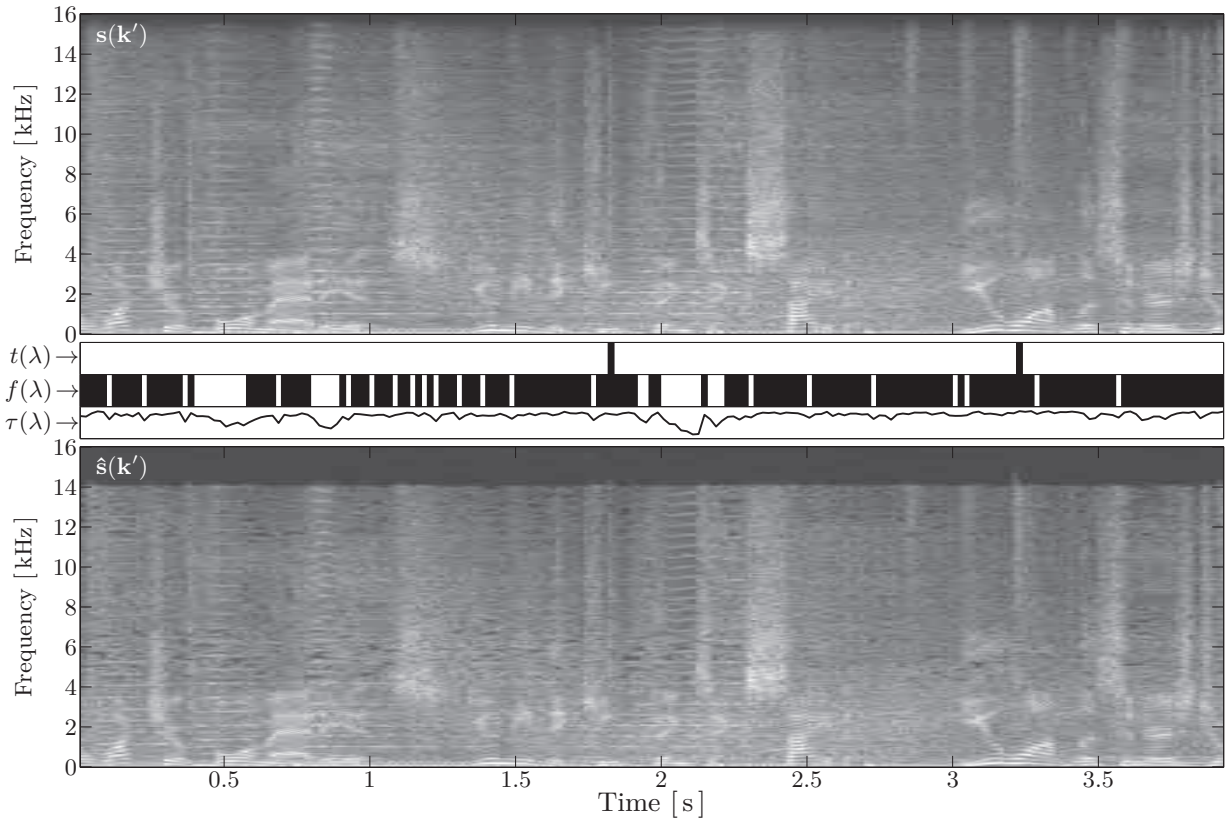
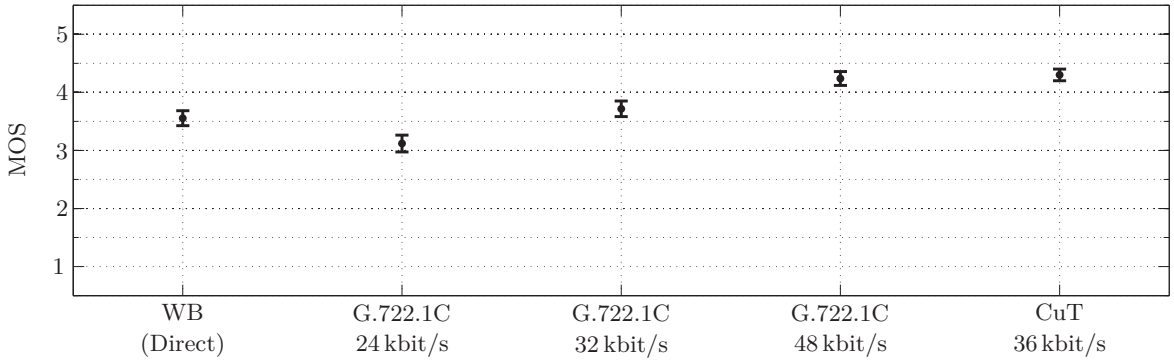


Figure 3.23: Example spectrograms for the super-wideband bandwidth extension algorithm (G.729.1-SWB Candidate B, bit rate: 36 kbit/s). Top: Input signal, Bottom: Encoded signal.

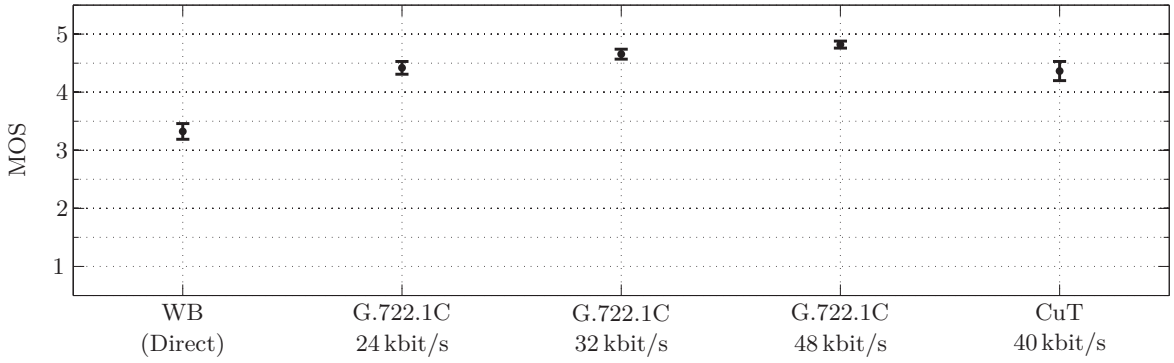
Example

A spectrogram of an example super-wideband signal $s(k')$ with mixed speech and audio content is shown in Figure 3.23 together with its transcoded (bandwidth extended) counterpart $\hat{s}(k')$. The wideband part $s_{\mathrm{wb}}(k)$ of the signal corresponds to the output of the G.729.1 baseband codec at a bit rate of 32 kbit/s. The extension band signal $\hat{s}_{\mathrm{uhb}}(k)$ is regenerated with 4 kbit/s of additional information. The figure also visualizes some results of the parametric signal analysis of $s_{\mathrm{uhb}}(k)$ according to Section 3.3.1, i.e., the parameters $t(\lambda)$, $f(\lambda)$, and $\hat{\tau}(\lambda)$.

The flag $f(\lambda)$ determines the *method* for spectral fine structure regeneration while the *tonality* of the regenerated extension band signal follows the (quantized) inverse tonality value $\hat{\tau}(\lambda)$. In the example, spectral replication ($f(\lambda) = 1$, black bar) is predominantly used, which is typical for speech signals and for noise-like extension band components. However, there are a few periods within the signal where harmonic synthesis ($f(\lambda) = 1$, white bar) is applied (e.g., 0.4 – 0.6 s). The harmonic components in the extension band signal are cleanly reproduced. Concerning the temporal signal envelope, $s_{\mathrm{uhb}}(k)$ is mostly classified as "stationary" ($t(\lambda) = 0$). Only two transient frames ($t(\lambda) = 1$, time indices 1.8 s and 3.2 s) are detected. The corresponding sharp onsets are accurately reproduced in the synthetic signal $\hat{s}_{\mathrm{uhb}}(k)$.



(a) Experiment 1a: mono clean speech, conducted at Huawei listening lab.



(b) Experiment 3a: mono music, conducted at Dynastat listening lab.

Figure 3.24: Subjective listening test results for the MDCT domain super-wideband bandwidth extension algorithm.

Subjective Test Results

The quality of the described bandwidth extension algorithm has been assessed in the course of the qualification phase of the ITU-T standardization process where a series of subjective listening tests has been conducted in order to compare the proposed "Codec under Test" (CuT) with the existing ITU-T super-wideband codec G.722.1 Annex C [ITU-T 1999, Xie et al. 2006, Lamblin et al. 2008]. The applied test methodology was the so called "triple stimulus/hidden reference/double blind" test method, in short "Ref-A-B" [ITU-R 1997]. In this method, the "reference" is the unprocessed signal and the samples A and B are, in random order, the test sample and again the (hidden) reference. "Ref-A-B" is designed as an *expert listener* method. Therefore, this test is able to assess the *listening expertise* of the subjects by evaluating their ability to identify the hidden reference. Hence, unreliable votes can be excluded from the evaluation. Finally, the votes of 24 subjects, ranging from 1.0 (very annoying impairment) to 5.0 (imperceptible impairment), have been accepted and used for each test condition. As an overall result, the codec passed *all* mono quality requirements as defined in the "Terms of Reference."

The test conditions which are particularly relevant for the present bandwidth extension algorithm have been evaluated in Experiment 1a (clean speech) and in

Experiment 3a (mixed content / music). The respective quality scores are reproduced in Figure 3.24, showing the mean scores and the associated 95% confidence intervals.

For clean *speech*, the proposed codec is clearly better than the G.722.1C codec at similar bit rates. Its performance (at 36 kbit/s) is in fact comparable to that of G.722.1C at 48 kbit/s. The effective bit rate saving of 12 kbit/s can be mainly attributed to the proposed novel bandwidth extension techniques.

For mixed content and *music* signals, the codec has been used at a bit rate of 40 kbit/s. Here, the quality is slightly inferior to the (non-parametric) transform codec G.722.1C. Nevertheless, it could still be shown that the parametric bandwidth extension algorithm can deliver a major quality gain compared to the *original* (direct) wideband signal (Δ-MOS = 1.04). Therefore, it cannot be ruled out that the small quality loss compared to G.722.1C is due to the *wideband* part of the codec (ITU-T G.729.1).

In summary, the proposed codec can (unlike G.722.1C at 24 and 32 kbit/s) offer a stable quality for varying sources, i.e., for *both* speech and music. In Experiment 2a of the test series, this finding could also be confirmed for *noisy speech* (mean subjective score with office noise: 4.76, music noise: 4.54). However, the tested coding mode in this experiment (48 kbit/s) is predominantly non-parametric.

More subjective listening test results for the MDCT domain bandwidth extension algorithm are presented in Chapter 6 where the concept of embedded coding is compared with parameter estimation (Chapter 4) and hidden transmission techniques (Chapter 5).

Objective Test Results

For a more detailed analysis of *individual* algorithmic components, objective quality measurements are very useful. Here, the quality impact of the proposed adaptive temporal envelope (ATE) control in conjunction with the integrated frame erasure concealment (FEC) mechanism (cf. Figure 3.20) shall be objectively assessed.

Table 3.6 lists the measured quality gains in terms of a PEAQ score improvement [ITU-R 1998] compared to a codec version where temporal normalization (3.23) and denormalization (3.27) have been disabled. For the test, the parametric coding mode (32+4 kbit/s) of the proposed codec has been used. The four test items have been taken from the EBU SQAM corpus [EBU 1988]. For reference, the average number of bits per frame which is used for the description of the ATE and FEC information is tabulated.[6]

[6] Note that, if the FEC functionality is not desired, 6 (or 1) redundant bits can be saved for each stationary (or transient) frame without compromising the performance under clean channel conditions, cf. Section 3.3.2.

Table 3.6: PEAQ improvements (Δ-PEAQ) by applying the temporal gain function $\hat{g}_{\mathrm{TGF}}(k)$. Codec mode: 36 kbit/s.

Test Item	Average number of bits per frame	0% FER	5% FER	10% FER	10% FER (SWB only)
Castanets	14.49	+0.17	+0.14	+0.11	+0.20
German Male	13.58	+0.13	+0.04	+0.04	+0.20
German Female	14.33	+0.27	+0.05	+0.06	+0.27
Pop Music (ABBA)	13.34	+0.33	+0.12	+0.05	+0.31
∅	13.94	+0.23	+0.09	+0.07	+0.25

Surprisingly, it can be observed that the achieved quality gain decreases with increasing frame erasure rate. This behavior can be explained by the overall dominance of the (unavoidable) errors that are introduced by the low band (0 – 8 kHz) FEC module of the codec [Vaillancourt et al. 2007]. This proposition is verified by limiting the frame erasures to the bits related to bandwidth extension (8 – 14 kHz). Therefore, "partial" erasures have been introduced at the same (pseudo-random) bitstream positions as before. This way, the performance gain through the proposed FEC module can be measured independently from the 0 – 8 kHz FEC. The results are shown in the last column of the table which lists Δ-PEAQ values for a bit rate of 36 kbit/s and 10% SWB erasure. Obviously, the initial quality gain at 0% FER (+0.23) can be maintained even at 10% SWB erasure (+0.25). The new FEC algorithm is therefore highly effective, i.e., the codec can be expected to operate well even under difficult network traffic conditions.

It should be noted that the present FEC mechanism can also be used to elegantly conceal short-term *bandwidth switchings* that occur when the bit rate of the embedded codec is lowered from 36 kbit/s down to 32 kbit/s for a short period, e.g., due to intermittent network congestion. In contrast, in the standardized version of the super-wideband codec [Laaksonen et al. 2010], such situations are handled by *estimating* the missing high frequency band from the received wideband signal using artificial bandwidth extension techniques, i.e., without the help of auxiliary information, cf. Chapter 4.

Algorithmic Complexity

The algorithmic complexity of the super-wideband bandwidth extension algorithm has been measured by instrumenting the source code according to [ITU-T 1993*a*]. The complexity is given in weighted million floating point operations per second (WMOPS) for the *worst case* that was observed among all encoded signal frames. Thereby, the complexity of the G.729.1 baseband codec (36 WMOPS, fixed point) is excluded from the measurement.

For mono input at 32 kHz sampling rate and at a bit rate of 36 kbit/s, the encoder part of the bandwidth extension algorithm, i.e., the parameter analysis (Sec-

tion 3.3.1) and the quantization (Section 3.3.2) modules, requires 5.03 WMOPS. The decoder complexity is 5.27 WMOPS. For a bit rate of 40 kbit/s, the complexity is slightly higher, i.e., 5.42 WMOPS for the encoder side and 5.57 WMOPS for the decoder side.

These complexity figures are obviously higher than for the TDBWE algorithm (see Table 3.3). This can, first, be explained by the higher sampling rate of 16 kHz and, second, by the inclusion of the QMF filterbank and of the preprocessing operations in the present complexity measurements. Naturally, also the increased demands of the codec design, i.e., suitability for audio signals, lead to a certain complexity increase.

Algorithmic Delay

The bandwidth extension algorithm does not increase the delay of the baseband codec significantly because most of the parametric analysis and synthesis techniques are conducted in the MDCT domain (cf. Section 3.3.1). The temporal envelope analysis does not use any look-ahead samples and the temporal gain function is constructed without any knowledge of the succeeding frame. Hence, there is only one source of additional delay in the codec, i.e., the QMF filterbank, see Figure 3.21 and Figure 3.22. Yet, owing to the IIR solution with phase equalization (Section 2.1.2, [Löllmann et al. 2009]), the algorithmic delay of G.729.1 (48.9375 ms) is only increased by 2.21875 ms which is considerably lower than for a competitive FIR QMF solution. If, in addition, the G.729.1 FIR QMF bank is replaced by the IIR solution, the total delay increase is merely 0.72 ms.

3.3.7 Discussion

The described, novel super-wideband bandwidth extension algorithm, as a part of the ITU-T candidate codec, succeeded in fulfilling the official requirements for mono input signals. The test results confirm that the codec can offer high performance with comparatively low computational complexity. In particular the low bit rates of 36 and 40 kbit/s provide compact but comprehensive information to synthesize additional frequency content.

As one more fundamental result, it could be shown that a concise parametric signal analysis and synthesis is feasible directly in the MDCT domain despite certain shortcomings of this representation (see, e.g., Figure 3.16). In other codecs, typically, a *second* frequency transform (and inverse transform) is implemented for parametric analysis and synthesis, see, e.g., [Dietz et al. 2002]. A benefit of the proposed solution is that the implementation in the MDCT domain facilitates a tight integration with subsequent transform coding stages. For instance, the spectral envelope parameters are reused as scale factors for spherical vector quantization of MDCT coefficients.

In comparison to the previously discussed TDBWE algorithm (Section 3.2), the following advantages and differences can be identified:

- Instead of the simple (and quite rigid) temporal envelope representation of the TDBWE algorithm, a new, adaptive method is used where the temporal resolution is adapted to the characteristics of the current input frame.

- A novel, highly effective method for frame erasure concealment has been directly incorporated into the temporal envelope shaping module of the bandwidth extension scheme. In contrast, for the TDBWE algorithm, a plain parameter repetition approach is taken in case of a lost frame.

- In the TDBWE case, the spectral fine structure was regenerated based on an elaborate model of speech production (Section 3.2.3). The present algorithm cannot rely on a specific source model, and therefore more generic techniques are applied, namely spectral replication and harmonic synthesis, where, in both cases, the degree of tonality is adapted to match the original input signal. In particular, special care needs to be taken in the harmonic analysis and synthesis in the MDCT domain which, as such, is not well suited for this application.

- Both the TDBWE algorithm and the harmonic synthesis method allow to represent the spectral fine structure as a mixture of noise and harmonic signals. However, it could be observed that the super-wideband bandwidth extension algorithm is even capable of reproducing *disharmonic* signals (e.g., triangle) reasonably well. The spectral fine structure of such signals is regenerated by producing the "closest matching" harmonic signal and by suppressing unwanted sinusoids with the (relatively fine-grained) spectral envelope.

Summarizing, an increased flexibility concerning the type of input signals (speech, music, mixed, ...) is offered by the new super-wideband bandwidth extension algorithm in comparison to the TDBWE approach. Naturally, this is in part due to the fact that the human auditory system is even more insensitive to the fine acoustic details at frequencies beyond 8 kHz than to the details of the 4 – 8 kHz frequency range. Still, the main reason for the enhanced performance are the improved modeling capabilities and the adaptive nature of the employed signal parametrization which facilitates a more flexible and therefore better signal representation and synthesis.

The enhanced capabilities of the algorithm, naturally, come at the cost of a higher complexity, a factor of approximately three in comparison to TDBWE. However, a factor of two can already be attributed to the doubling of the sampling frequency in the extension band.

3.4 Comparison with Other Approaches

Meanwhile, the integration of bandwidth extension techniques into speech and audio codecs can look back on a long history, beginning with early ideas such as proposed in [Un & Magill 1975] or [Makhoul & Berouti 1979]. Here, a comprehensive survey of the related proposals that have been made in the literature shall be provided.

Bandwidth Extension in the Time Domain

A first *standardized* speech codec which did not explicitly encode the higher frequency portion of the speech signal is the GSM FullRate codec [ETSI 1990, Vary et al. 1988], which is based on the RPE-LTP principle (Regular Pulse Excitation with Long-Term Prediction). In this codec, spectral replica of a baseband excitation signal are (implicitly) used as higher band excitation. In the "Pyramid CELP" codec of [Erdmann et al. 2002, Erdmann 2005], the spectral replica are successively replaced by the actual encoded LP residual in the respective frequency bands, thus forming a hierarchical codec with several bitstream layers.

Instead, [Paulus & Schnitzler 1996, Paulus 1997] *explicitly* use parametric bandwidth extension techniques to extend a CELP speech codec. The parameter set to describe the 6 – 7 kHz extension band signal consists of 5 ms-subframe gains. As a further improvement, the spectral envelope in the extension band can be extrapolated by using a "shadow codebook," cf. [Schnitzler 1998, Schnitzler 1999]. These investigations finally led to a candidate proposal for the 3GPP AMR-WB codec, see [Erdmann et al. 2001]. Also the candidate codec which is described in [McCree 2000, McCree et al. 2001] uses a parametric bandwidth extension approach. Here, the spectral envelope for the 4 – 7 kHz extension band is explicitly transmitted while the temporal envelope is extrapolated from the 3 – 4 kHz range. In the final AMR-WB standard [ETSI 2001*b*, Bessette et al. 2002], the frequency band from 6.4 kHz to 7 kHz is artificially generated and spectrally shaped with the help of an LPC filter with extrapolated coefficients. Optionally, a gain correction factor can be transmitted (only in the 23.85 kbit/s mode of the codec). This method is also applied in the codecs that are derived from AMR-WB, namely 3GPP2 VMR-WB [3GPP2 2005, Jelínek et al. 2004] and ITU-T Rec. G.718 [ITU-T 2008*a*, Vaillancourt et al. 2008]. Two other related proposals from the literature are [Taori et al. 2000] and [Valin & Lefebvre 2000].

The relatively simple time domain approach of the AMR-WB codec has been extended in the AMR-WB+ codec [ETSI 2004*a*, Makinen et al. 2005] to encode the upper half of the input spectrum. A spectral envelope (8 LPC coefficients, quantized in the LSF domain) and a temporal envelope (4 gain correction factors) are explicitly transmitted using 16 bit per frame (256 samples). The actual bit rate depends on the selected sampling frequency. The AMR-WB+ decoder applies the temporal envelope to a spectral copy of the baseband signal. Then, after a "buzziness reduction" processing stage, the spectral envelope is restored using LPC

synthesis. Finally, "envelope smoothing" is applied to the extension band as a final post-processing step.

Another bandwidth extension algorithm operating in the time domain is standardized with the 3GPP2 EVRC-WB codec [3GPP2 2010, Krishnan et al. 2007]. In the encoder, a first preprocessing step removes intermittent transients or "clicks" from the extension band signal. This measure is supposed to be perceptually (nearly) transparent while easing parametric encoding (e.g., a coarser representation of the temporal envelope can be used). The parameter set for a 20 ms frame consists of 6 LPC coefficients, an explicitly encoded harmonic-to-noise ratio, and 5 gain correction factors (4 ms subframes). The latter are obtained in a closed-loop approach by providing the decoder-side synthetic extension band signal also in the encoder. The total bit rate for bandwidth extension is 0.8 kbit/s in voiced frames and 1.35 kbit/s in unvoiced frames. The decoder regenerates the high frequencies by applying a non-linearity and spectral whitening to the baseband signal. To correct the tonality, temporally modulated noise is added according to the transmitted harmonic-to-noise ratio. The generated signal is filtered through the LPC synthesis filter. Finally, the correction gains are applied.

Bandwidth Extension in the Frequency Domain

Meanwhile, a number of codecs and codec proposals exist that perform a parametric bandwidth extension in the *frequency domain*. For example the audio coding tools of the MPEG-4 standard [ISO 2005] use a technique labeled "Spectral Band Replication" (SBR) [Ekstrand 2002, Dietz et al. 2002]. A 64-channel complex-valued Pseudo-QMF filterbank is used to represent the audio signal. Several parameters are extracted from this representation, namely a spectral envelope over time frames of different length (signal-adaptive), a tonal-to-noise ratio, and a number of single sinusoids which are not present in the replicated spectral details. The decoder, first, analyzes the baseband signal with the Pseudo-QMF filterbank to provide a basis for the spectral replication approach. Filtered subbands of the baseband are used as extension band excitation signal. Then, the adaptive spectral envelope is applied and the sinusoids as well as a controlled amount of noise are added. For generic audio signals, average bit rates well below 4 kbit/s (depending on the sampling rate, cf. [Larsen & Aarts 2004, Chapter 5]) can be achieved with the SBR technique.

SBR has been extensively used in several codec standards, beginning with extensions to MPEG1-Layer-2 [Schug et al. 2003] and Layer-3 [Ziegler et al. 2002] (mp3PRO). It is also part of the 3GPP "enhanced aacPlus" codec [ETSI 2004*b*, Ehret et al. 2003]. The corresponding variant of the MPEG AAC codec which incorporates SBR is called "high efficiency AAC" (HE-AAC) [Wolters et al. 2003]. A *modified* version of the SBR algorithm is used in the AAC-ELD codec [Schnell et al. 2007, Schnell et al. 2008] to achieve a reduced overall algorithmic delay of 42 ms. Finally, an enhanced SBR scheme (eSBR) is being considered for the "Uni-

fied Speech and Audio Codec" (USAC) being developed by the MPEG committee, see [Neuendorf et al. 2009, Neuendorf et al. 2009a]. Among other modifications, new methods for the replication of the spectral details are under investigation [Nagel & Disch 2009, Nagel et al. 2010].

The SBR technique must be applied as a "wrapper" around a given codec because the employed frequency transform (Pseudo-QMF) is usually incompatible with the core codec transform. Also, the decoded baseband signal has to be analyzed again within the decoder. Hence, computational complexity as well as algorithmic delay are increased. Therefore, in the scope of several other standardization approaches, a trend towards a more tight integration with the core transform (mostly the MDCT, cf. Section 3.3) can be observed.

In particular, Amd. 6 to ITU-T Rec. G.729.1 as well as Amd. 2 to ITU-T Rec. G.718 [ITU-T 2006, ITU-T 2008*a*, Laaksonen et al. 2010] use parametric MDCT domain bandwidth extension techniques. At a codec rate of 36 kbit/s, a 4 kbit/s bitstream layer is used for this purpose. The transmitted parameter set depends on a decision (1 bit) between a "generic" and a "sinusoidal" coding mode. In the "generic" mode, used in non-tonal frames, four frequency indices are transmitted to achieve an adaptive spectral patching (see Section 2.5.1). These four frequency bands are scaled in a two-step procedure using spectral gain parameters. Additionally, two so called "sinusoidal components" are added to the MDCT spectrum. In fact, these "sinusoids" are merely quantized amplitudes of the MDCT bins that exhibit the highest magnitude error. In particular, they do *not* represent sinusoidal signals in the time domain, see Figure 3.16 and [Daudet & Sandler 2004]. In tonal frames, i.e., in the "sinusoidal" coding mode, *ten* individual "sinusoidal components" are transmitted. The decoder implements a blind method for pre-echo reduction as well as a simple FEC scheme based on spectrum repetition and attenuation. In Chapter 6, this codec (ITU-T Rec. G.729.1 Amd. 6 at 36 kbit/s) is used as a super-wideband reference codec. It is also compared with the candidate codec as described in Section 3.3.

Another bandwidth extension approach in the MDCT domain has been standardized in ITU-T Rec. G.719 [ITU-T 2008*b*, Xie et al. 2009]. Here, the spectral envelope is jointly encoded with the baseband envelope. The bandwidth extension is based on spectral folding techniques whereby the transition frequency between baseband and extension band components is signal-adaptive. Additionally, there may also be frequency bands within the baseband signal that did not receive any bit allocation. In these bands, "noise filling" is applied which can be seen as a form of bandwidth extension within these narrow passbands. A mixture of noise filling and spectral folding is also used in the CELT codec [Valin et al. 2010].

Some other examples from the literature that implement bandwidth extension in the MDCT domain include [Oshikiri et al. 2002, Oshikiri et al. 2004, Oshikiri et al. 2007] for wideband and super-wideband scenarios.

Conclusions

While the initial motivation for the integration of bandwidth extension techniques into time and transform domain codecs was increased coding efficiency as well as decreased complexity, another major advantage is the possibility to enhance widely deployed coding standards with additional bitstream layers while preserving interoperability with existing infrastructure and equipment.

Typical bit rates for parametric bandwidth extension range from 0.5 up to 4 kbit/s depending on the application scenario. Meanwhile, even lower rates have been reported to deliver adequate quality, at least for speech signals. The respective coding techniques often use *predictive* quantization schemes that explicitly exploit baseband information, e.g., [Ehara et al. 2005, Agiomyrgiannakis & Stylianou 2004, Agiomyrgiannakis & Stylianou 2007, Geiser & Vary 2007*a*]. The case of *zero* additional bit rate, i.e., a purely *artificial* bandwidth extension approach will be discussed in the following chapter.

As shown by the above literature survey, parametric bandwidth extension and the integration with a baseband codec have become very popular in recent years. Other reviews of the pertinent literature can, e.g., be found in [Larsen & Aarts 2004, Chapter 5] and in [Geiser, Ragot & Taddei 2008]. However, despite the extensive literature on the topic, for the concrete algorithms that have been proposed and discussed in this thesis, still a number of new ideas and approaches were required to arrive at solutions that actually fulfill *all* requirements and demands of the respective application scenarios.

Chapter 4

Receiver Based Parameter Estimation

The transmission of high quality speech and audio signals with a cutoff frequency of 7 kHz or higher is generally deemed a highly desirable feature for the telephone network and future audio communication systems. However, the required costly and time-consuming modifications of network equipment and the related communication protocols turned out to be a major obstacle for the introduction of (long existing) high quality speech and audio coding techniques in today's networks.

An alternative, promising approach to improve the quality of the received band-limited speech signals is the deployment of *artificial* bandwidth extension (ABWE) techniques, where the limited frequency range of narrowband speech is *artificially* extended at the *receiving* end. This approach has, e.g., been studied in [Carl & Heute 1994, Jax 2002, Kornagel 2006, Pulakka & Alku 2011]. The related techniques might, as anticipated in [Jax & Vary 2006], be able to speed up the narrow- to wideband change-over of communication networks.

The goal of this chapter is to study the *receiver based estimation* of the bandwidth extension parameter sets that have been introduced in Chapter 3 for embedded coding. To allow a direct comparison of the achieved speech quality of both approaches, a subjective listening test has been conducted. The respective details and the test results are presented in Chapter 6.

4.1 Overview

The block diagram in Figure 4.1 illustrates the principle of the ABWE algorithm that is used in this chapter. First, a so called *feature vector* $\mathbf{x}_\mathrm{f}(\lambda)$ is derived from the decoded baseband signal $\hat{s}_\mathrm{bb}(k)$ or, if possible, directly from the bitstream of the baseband codec. This feature vector is supposed to compactly describe the relevant characteristics of the current speech frame. Second, with the help of a pre-trained statistical model, the parameter set $\hat{\mathbf{p}}(\lambda)$ for bandwidth extension is estimated from the current (and previous) feature vectors. The remaining components of the block diagram, i.e., the extension band synthesis and the final synthesis filterbank, have already been described in the Chapters 2 and 3.

An important requirement for a concise statistical estimation of the parameter vector $\mathbf{p}(\lambda)$ from the observed feature vector $\mathbf{x}_\mathrm{f}(\lambda)$ is a sufficient amount of *mutual information* $I(\mathbf{x}_\mathrm{f};\mathbf{p})$ [Cover & Thomas 1991] that is shared between these two variables. It is therefore important to note that ABWE as such is only applicable

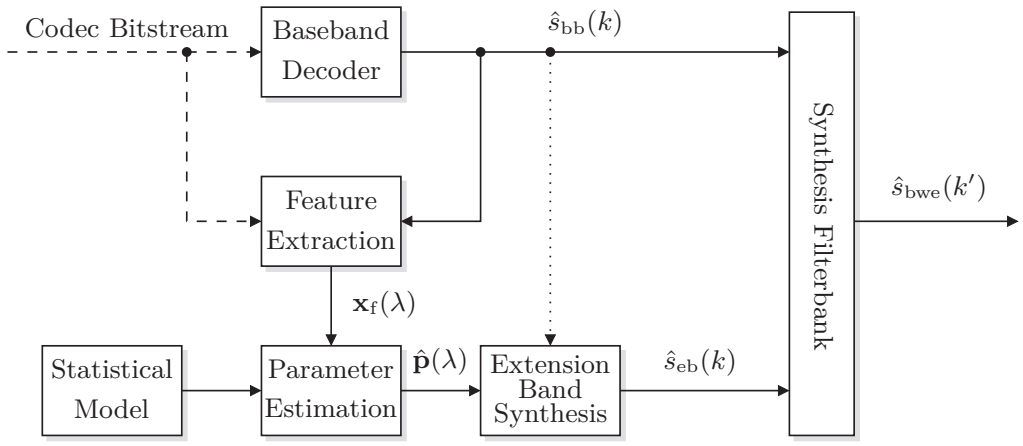


Figure 4.1: System for artificial bandwidth extension (ABWE) of band-limited speech signals.

to *speech* signals because, in contrast to general audio signals, a well-understood *source model* is available for this signal class. The different frequency bands of speech signals are actually produced by the same physical sound source and certain statistical dependencies can therefore be expected. Unfortunately, the actual amount of mutual information that is shared between the baseband and the extension band is relatively low. The practically achievable performance of ABWE algorithms is therefore limited. Yet, a certain, consistent quality improvement has been reported in various studies and publications. Particular measurements of mutual information $I(\mathbf{x}_{\mathrm{f}}; \mathbf{p})$ for the narrowband-to-wideband bandwidth extension scenario are available in [Nilsson et al. 2002, Jax & Vary 2002, Jax & Vary 2004, Nour-Eldin et al. 2006], and [Geiser et al. 2007].

A particular statistical estimation method, based on Hidden Markov modeling of the extension band parameters, has been described in [Jax 2002] and [Jax & Vary 2003]. It is summarized in Section 4.2. This method shall be applied to the bandwidth extension parameter sets and algorithms as described in the previous chapter. Concretely, two example realizations for ABWE systems, i.e., for wideband and super-wideband speech signals, are considered:

- In Section 4.3, the estimation of the standardized TDBWE (Time Domain Bandwidth Extension) parameter set for narrowband-to-wideband bandwidth extension, cf. Section 3.2 and [Geiser et al. 2007a], is described. In a first implementation, the TDBWE parameters are estimated within the context of the standardized ITU-T G.729.1 codec which has been designed for VoIP applications. In an alternative implementation, the TDBWE algorithm is used in combination with the 3GPP EFR codec which is the predominant codec in today's mobile telephone networks.

- In Section 4.4, a wideband-to-super-wideband ABWE scenario is considered, i.e., the super-wideband parameter set of Section 3.3 is estimated based on the received wideband signal.

After a brief discussion (Section 4.5), the chapter concludes with a survey of other approaches for artificial bandwidth extension that have been described in the literature (Section 4.6).

4.2 Theoretical Background

This section summarizes the design of minimum mean square error (MMSE) estimators for the parameter vectors $\mathbf{F}$ (frequency envelope) and $\mathbf{T}$ (temporal envelope) of the previous chapters. These estimators are based on the Hidden Markov modeling (HMM) framework of [Jax 2002, Jax & Vary 2003].

4.2.1 Features

As a prerequisite, a sequence of *feature vectors* $\mathbf{x}_\mathrm{f}(\lambda) \in \mathbb{R}^b$, which compactly describes the received baseband signal $\hat{s}_\mathrm{bb}(k)$, is defined:

$$\mathbf{X}_\mathrm{f} \doteq \{\mathbf{x}_\mathrm{f}(1), \ldots, \mathbf{x}_\mathrm{f}(\lambda)\}, \tag{4.1}$$

where the numbers 1 to λ specify the respective frame indices. In particular, λ designates the *current* frame, i.e., $\mathbf{x}_\mathrm{f} = \mathbf{x}_\mathrm{f}(\lambda)$ in the following. The actual feature selection depends on the specific application. This is discussed in Section 4.3.2 and in Section 4.4.2 for the wideband and the super-wideband case, respectively.

It should be noted that it is beneficial to use *mutually decorrelated* features, which facilitates a simplified modeling of the related probability densities, see (4.7). The feature decorrelation can, e.g., be enforced with a *Karhunen-Loève Transform* (KLT), i.e., multiplication of $\mathbf{x}_\mathrm{f}$ with the matrix of eigenvectors of its covariance matrix.

4.2.2 Derivation of the MMSE Estimation Rule

For the present description, the "generic" estimation quantity $\mathbf{p}$ is used instead of the actual parameter vectors $\mathbf{F}$ and $\mathbf{T}$. The criterion for MMSE estimation of a vector $\mathbf{p}$ with given observations $\mathbf{X}_\mathrm{f}$ is

$$E\left\{||\mathbf{p} - \hat{\mathbf{p}}||^2 \,\middle|\, \mathbf{X}_\mathrm{f}\right\} \rightarrow \min, \tag{4.2}$$

where $\hat{\mathbf{p}}$ is the estimation result. The solution to this optimization problem is the conditional expectation, cf. [Kay 1993, Jax & Vary 2003]:

$$\hat{\mathbf{p}}_\mathrm{MMSE} = E\left\{\mathbf{p} \,\middle|\, \mathbf{X}_\mathrm{f}\right\}. \tag{4.3}$$

Using a pre-computed vector codebook $\mathcal{C} = \{\hat{\mathbf{p}}_0, \ldots, \hat{\mathbf{p}}_{N_\mathcal{C}-1}\}$ for the vectors $\mathbf{p}$ (e.g., obtained with the LBG algorithm [Linde et al. 1980]), this MMSE estimate can be expressed as

$$\hat{\mathbf{p}}_\mathrm{MMSE} = \sum_{\hat{\mathbf{p}}_i \in \mathcal{C}} \hat{\mathbf{p}}_i \cdot P(\hat{\mathbf{p}}_i \,|\, \mathbf{X}_\mathrm{f}) \tag{4.4}$$

which essentially is a weighted sum over the $N_\mathcal{C}$ centroids of the codebook $\mathcal{C}$. Thereby, the weights $P(\hat{\mathbf{p}}_i \,|\, \mathbf{X}_\mathrm{f})$ designate *a posteriori probabilities*.

4.2.3 A Posteriori Probabilities

The probabilities $P(\hat{\mathbf{p}}_i \,\big|\, \mathbf{X}_\mathrm{f})$ from (4.4) can be reformulated as

$$P(\hat{\mathbf{p}}_i \,\big|\, \mathbf{X}_\mathrm{f}) = \frac{p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f})}{p(\mathbf{X}_\mathrm{f})} = \frac{p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f})}{\sum\limits_{\hat{\mathbf{p}}_i \in \mathcal{C}} p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f})}. \tag{4.5}$$

The joint densities $p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f})$ in (4.5) are computed as the product of the so-called *observation densities* $p(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_i)$ of the *current* feature vector $\mathbf{x}_\mathrm{f}$ and of specific joint densities that comprise accumulated *a priori knowledge*:

$$p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f}) = p(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_i) \cdot p(\hat{\mathbf{p}}_i, \mathbf{X}'_\mathrm{f}), \tag{4.6}$$

where $\mathbf{X}'_\mathrm{f} \doteq \{\mathbf{x}_\mathrm{f}(1), \ldots, \mathbf{x}_\mathrm{f}(\lambda - 1)\} = \mathbf{X}_\mathrm{f} \setminus \mathbf{x}_\mathrm{f}(\lambda)$.

The *observation densities* $p(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_i)$, i.e., the first factor of (4.6), are approximated with *Gaussian Mixture Models* (GMMs) with L_GMM mixture components:

$$p(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_i) \approx \sum_{l=0}^{L_\mathrm{GMM}-1} \rho_{il} \cdot \frac{\mathrm{e}^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{il})^\mathrm{T} \mathbf{V}_{il}^{-1} (\mathbf{x}-\boldsymbol{\mu}_{il})}}{\sqrt{(2\pi)^b \cdot \det \mathbf{V}_{il}}}, \tag{4.7}$$

whereby the mean vectors $\boldsymbol{\mu}_{il}$, the covariance matrices $\mathbf{V}_{il}$, as well as the mixture weights ρ_{il} are determined during an offline training phase, cf. [Moon 1996]. If the features $\mathbf{x}_\mathrm{f}$ are decorrelated, e.g., using a KLT, the matrices $\mathbf{V}_{il}$ can be assumed to be diagonal. This restriction leads to a significantly reduced complexity.

The second factor of (4.6), i.e., the joint probability density function $\alpha_i(\lambda) \doteq p(\hat{\mathbf{p}}_i, \mathbf{X}'_\mathrm{f}) = p(\hat{\mathbf{p}}_i, \mathbf{x}_\mathrm{f}(1), \ldots, \mathbf{x}_\mathrm{f}(\lambda - 1))$ can be computed recursively, i.e.,

$$\alpha_i(1) = P(\hat{\mathbf{p}}_i) \tag{4.8}$$

and

$$\alpha_i(\lambda + 1) = \sum_{j=0}^{N_\mathcal{C}-1} \alpha_j(\lambda) \cdot p\left(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_j(\lambda)\right) \cdot P\left(\hat{\mathbf{p}}_i(\lambda + 1) \,\big|\, \hat{\mathbf{p}}_j(\lambda)\right), \tag{4.9}$$

whereby the discrete probabilities $P(\hat{\mathbf{p}}_i)$ and $P\left(\hat{\mathbf{p}}_i(\lambda + 1) \,\big|\, \hat{\mathbf{p}}_j(\lambda)\right)$ have to be predetermined as "a priori knowledge" during the offline training phase. This method to compute the a priori term $p(\hat{\mathbf{p}}_i, \mathbf{X}'_\mathrm{f})$ explicitly exploits a priori knowledge of *first* order, i.e., the *state transitions* of the first order Hidden Markov Model are explicitly considered therein.

4.2.4 MMSE Estimation

Given an actual observation $\mathbf{x}_\mathrm{f}$, the trained GMMs (4.7) are used to compute $p(\mathbf{x}_\mathrm{f} \,\big|\, \hat{\mathbf{p}}_i)$ for all $i \in \{0, \ldots, N_\mathcal{C} - 1\}$. These values are used to update the a priori term according to (4.9). Then, the joint density $p(\hat{\mathbf{p}}_i, \mathbf{X}_\mathrm{f})$ is computed by multiplying both terms according to (4.6). Finally, insertion of (4.6) into (4.5) and then (4.4) yields the desired MMSE estimate $\hat{\mathbf{p}}_\mathrm{MMSE}$ for the vector $\mathbf{p}$.

4.3 Estimation of TDBWE Parameters

The previously described statistical estimation scheme can be used to estimate the TDBWE parameter set for narrowband-to-wideband bandwidth extension which has been introduced in Section 3.2.1. The parameter estimation shall be primarily conducted within the framework of the ITU-T Rec. G.729.1 codec (Sections 4.3.1 – 4.3.5), cf. [Geiser et al. 2007]. However, the TDBWE algorithm can also be used to extend the bandwidth of the 3GPP EFR codec. This alternative implementation is briefly discussed in Section 4.3.6.

The specific setup for artificial bandwidth extension which is considered here allows to exploit various synergy effects. In particular, for G.729.1, a given implementation of the TDBWE algorithm, e.g., with optimized fixed point arithmetics, can be directly reused. Furthermore, the decoder does not only provide the synthesized narrowband speech signal but also the *parameters* from the baseband bitstream. In fact, these parameters may directly function as features in the estimation process; a similar approach is taken in [Deshpande & Ramakrishnan 2005]. Recognizing the described synergies, a very efficient solution for artificial bandwidth extension on top of the narrowband codec can be realized.

4.3.1 Reduced Parameter Set

In principle, it is feasible to estimate the entire TDBWE parameter set, i.e., the time envelope parameters $\mathbf{T}$ (3.1) and the frequency envelope parameters $\mathbf{F}$ (3.5) directly represent the vector $\mathbf{p}(\lambda)$ in the estimation process. However, for reasons of increased efficiency, the dimension of the parameter set may be reduced. It has been found that such dimensionality reduction does not necessarily degrade the achievable quality significantly, in particular if the estimated parameters are smoothed in a post-processing step (Section 4.3.4). For the TDBWE case, for instance, an *averaged* temporal envelope,

$$\bar{T}(\lambda) = \frac{1}{N_{\mathrm{TE}}} \cdot \sum_{\lambda_{\mathrm{SF}}=0}^{N_{\mathrm{TE}}-1} T(\lambda, \lambda_{\mathrm{SF}}), \tag{4.10}$$

i.e., a single gain can be used per frame. The (reduced) vector $\mathbf{p}(\lambda)$ is then used for model training. At the decoder side, the components of the estimated vector $\hat{\mathbf{p}}_{\mathrm{MMSE}}(\lambda)$ must be mapped back to the corresponding input parameters.

4.3.2 Narrowband Features

For the estimation of the vector $\mathbf{p}(\lambda)$, a relevant *feature vector* $\mathbf{x}_{\mathrm{f}}(\lambda)$ has to be chosen which shares sufficient mutual information with $\mathbf{p}(\lambda)$. Investigations about a proper choice for $\mathbf{x}_{\mathrm{f}}(\lambda)$ in ABWE systems have been published in [Jax & Vary 2004]: In terms of mutual information with the high band spectral envelope, the *autocorrelation coefficients* of the narrowband speech performed best, while in

terms of class separability, an advantage has been found for the *mel-frequency cepstral coefficients* (MFCCs). Yet, such features are not immediately available from the considered baseband codecs, i.e., their computation would consume additional complexity. A reasonable choice are the (also well performing) *line spectral pairs* (LSPs) from the CELP core layers that describe the narrowband spectral envelope. Additionally, the temporal envelope of the narrowband signal $\hat{s}_{\mathrm{nb}}(k)$ is included in the feature vector. This computation only requires little additional complexity. Summarizing, the feature vector $\mathbf{x}_{\mathrm{f}}(\lambda)$ is given by

$$\mathbf{x}_{\mathrm{f}}(\lambda) \doteq (\hat{\mathbf{q}}^{\mathrm{T}}(\lambda), \mathbf{T}_{\mathrm{nb}}^{\mathrm{T}}(\lambda))^{\mathrm{T}}. \tag{4.11}$$

where $\hat{\mathbf{q}}(\lambda)$ is the quantized LSP vector of the core layer CELP codec and $\mathbf{T}_{\mathrm{nb}}(\lambda)$ is the narrowband temporal envelope which is computed in the same manner as the high band time envelope in the TDBWE encoder.

4.3.3 Eligibility of the Feature Vector

The eligibility of the described feature vector $\mathbf{x}_{\mathrm{f}}$ for estimating the (reduced) TDBWE parameter set shall be assessed. For brevity, the following description only focuses on the frequency envelope parameter set $\mathbf{F}$, but it is nonetheless also valid for the temporal envelope parameters. Note that the frame index λ is also omitted for notational convenience.

A sufficient amount of mutual information between $\mathbf{x}_{\mathrm{f}}$ and the parameters $\mathbf{F}$ is a necessary condition for concise estimation results. Therefore, the mutual information $I(\mathbf{x}_{\mathrm{f}}; \mathbf{F})$ that is shared between the respective variables has been measured. Furthermore, as in [Nilsson et al. 2002], the *certainty* of $\mathbf{F}$ that is conveyed by the vector $\mathbf{x}_{\mathrm{f}}$ is computed as the ratio $I(\mathbf{x}_{\mathrm{f}}; \mathbf{F})/H(\hat{\mathbf{F}})$.[1] Thereby, $H(\hat{\mathbf{F}})$ constitutes the *entropy* of a uniformly *quantized* parameter vector. This quantization has to be introduced because the discrete Shannon entropy is not applicable to continuous quantities. In fact, $H(\hat{\mathbf{F}})$ is approximated by the *differential entropy* $h(\mathbf{F})$, see [Nilsson et al. 2002, Cover & Thomas 1991]:

$$H(\hat{\mathbf{F}}) \approx h(\mathbf{F}) - \log_2(\Delta^{\dim(\mathbf{F})}), \tag{4.12}$$

with the quantizer step size Δ. Here, Δ is set such that the resulting mean square error (MSE) $D = \dim(\mathbf{F}) \cdot \Delta^2/12$ is equal to the MSE of the TDBWE quantizer.

The actual measurements of $I(\mathbf{x}_{\mathrm{f}}; \mathbf{F})$ and $h(\mathbf{F})$ have been carried out using "k-nearest neighbor statistics" [Kraskov et al. 2004, Kozachenko & Leonenko 1987]. These methods are well suited to high-dimensional vector spaces while being data-efficient, i.e., only a comparatively small training set is required. The value for k has been selected as 1 in order to achieve minimum bias. Unfortunately, this choice of k inevitably increases statistical errors in the estimation of $I(\mathbf{x}_{\mathrm{f}}; \mathbf{F})$.

[1] The "certainty" $I(\mathbf{x}_{\mathrm{f}}; \mathbf{F})/H(\hat{\mathbf{F}}) \in [0, 1]$ designates the fraction of the information content of $\mathbf{F}$ that can be learned from the knowledge of the feature vector $\mathbf{x}_{\mathrm{f}}$.

Table 4.1: Measurements of mutual information $I(\mathbf{x}_\mathrm{f};\cdot)$, entropies $H(\cdot)$, and "high band certainty" $I(\mathbf{x}_\mathrm{f};\cdot)/H(\cdot)$. The given tolerances specify 95% confidence intervals.

$I(\mathbf{x}_\mathrm{f};\mathbf{F})$	$2.8111 \pm 0.0967\,\mathrm{bit/frame}$
$H(\hat{\mathbf{F}})$	$\approx 11.1670\,\mathrm{bit/frame}$
$I(\mathbf{x}_\mathrm{f};\mathbf{F})/H(\hat{\mathbf{F}})$	≈ 0.2517
$I(\mathbf{x}_\mathrm{f};\mathbf{T})$	$2.2836 \pm 0.1067\,\mathrm{bit/frame}$
$H(\hat{\mathbf{T}})$	$\approx 15.5619\,\mathrm{bit/frame}$
$I(\mathbf{x}_\mathrm{f};\mathbf{T})/H(\hat{\mathbf{T}})$	≈ 0.1467
$I(\mathbf{x}_\mathrm{f};\bar{T})$	$1.5719 \pm 0.0414\,\mathrm{bit/frame}$
$H(\hat{\bar{T}})$	$4.2999 \pm 0.0666\,\mathrm{bit/frame}$
$I(\mathbf{x}_\mathrm{f};\bar{T})/H(\hat{T})$	≈ 0.3656

Consequently, the experiments have been repeated several times with different speech samples and confidence intervals were computed for the obtained results. In particular, 120.000 active speech frames (20 ms length) from the NTT corpus [NTT 1994] were divided into 12 speech samples with 10.000 frames each.

The measurement results, both for the (reduced) time and for the frequency envelopes, are tabulated in Table 4.1. The values of $I(\mathbf{x}_\mathrm{f};\mathbf{F})$ and of the "certainty" $I(\mathbf{x}_\mathrm{f};\mathbf{F})/H(\hat{\mathbf{F}})$ give considerably larger numerical values than previously reported measurements [Jax & Vary 2004, Nilsson et al. 2002, Nour-Eldin et al. 2006]. Three reasons can be stated to support this finding:

- With a 4 kHz cutoff frequency instead of 3.4 kHz, the low frequency band is *wider* than in previous investigations, i.e., the obtained features are much more significant for artificial bandwidth extension.
- Conversely, the high band is *narrowed* from the range of $3.4 - 7$ kHz to $4 - 7$ kHz. This, of course, eases the estimation task.
- In previous investigations, a stricter distortion constraint of $D = 1$ dB has been used to estimate the entropy according to (4.12). Higher values for D decrease the estimated values of $H(\hat{\mathbf{F}})$ and therefore lead to higher certainty estimates. Though, the choice of D equal to the distortion of the standardized TDBWE quantizer (more than 3 dB) is actually justified since the TDBWE algorithm, as a part of ITU-T Rec. G.729.1, has been shown to provide a high speech quality.

Surprisingly, the certainty of the TDBWE time envelope $I(\mathbf{x}_\mathrm{f};\mathbf{T})/H(\hat{\mathbf{T}})$ is comparatively low. An explanation is its rather high temporal resolution with subframes of 1.25 ms which leads to increased entropy $H(\hat{\mathbf{T}})$. Since a smoothed time envelope is often sufficient for artificial bandwidth extension, the experiment has been

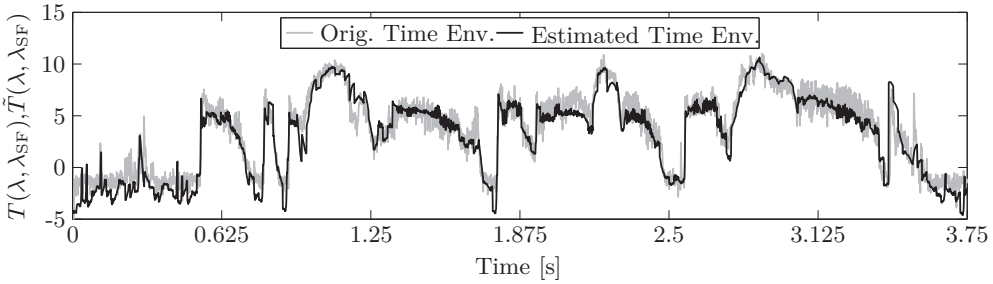


Figure 4.2: Example for the estimation of the time envelope.

repeated for an *averaged* time envelope $\bar{T}$ per frame (4.10), giving a significantly higher certainty value. From these results, it can be expected that the proposed scheme for G.729.1 TDBWE parameter estimation can indeed perform (at least) as well as other approaches for artificial bandwidth extension.

4.3.4 Parameter Post-Processing

The perceived wideband speech quality can be improved by smoothing the temporal evolution of the estimated parameters. As a very simple measure, a short gliding average can be applied to the estimated vectors, see also [Kornagel 2006]:

$$\hat{\mathbf{p}}(\lambda) = 0.5 \cdot (\hat{\mathbf{p}}_{\mathrm{MMSE}}(\lambda) + \hat{\mathbf{p}}_{\mathrm{MMSE}}(\lambda - 1)) \,. \tag{4.13}$$

This procedure attenuates artifacts that stem from strongly time-variant estimates.

4.3.5 Evaluation

The estimation performance that can be obtained with the described framework has been objectively measured based on 360 sec of clean speech from the NTT corpus. The MMSE estimators as described in Section 4.2 are based on 7 bit codebooks for $\mathbf{T}$ and $\mathbf{F}$ that have been trained via [Linde et al. 1980]. The GMMs for the observation densities $p(\mathbf{x}_{\mathrm{f}} \big| \hat{\mathbf{p}}_i)$ comprise eight mixture components each.

Estimation Performance

Figure 4.2 visualizes the performance of the parameter estimation based on a short utterance of a female English speaker. The original extension band time envelope is compared with the envelope contour that has been *estimated* from the baseband features. For the most part, the estimate follows the original envelope rather accurately, even for sounds of high energy (e.g., fricatives). Though, the estimate lacks some temporal fine structure which is coherent with the results from Section 4.3.3 (Table 4.1).

The *high band spectral distortion* is a common measure to assess the quality of bandwidth extension algorithms, e.g., [Jax 2002]. Here, 8th order AR models of the true and of the estimated high band signals[2] have been used to measure

[2]This model order has also been chosen in [Jax 2002] and [Jax & Vary 2003] (although for a wider extension band, i.e., for the frequency range of 3.4 – 7 kHz).

the spectral distortion. As a result, a high band *root mean square log spectral distortion* per frame of $\bar{d}_{\mathrm{LSD}} \approx 5.05\,\mathrm{dB}$ was measured for the artificially extended signal. Moreover, for reference, the corresponding value has also been computed for the *quantized* TDBWE output signal. The respective measurement yields $\bar{d}_{\mathrm{LSD,quant.}} \approx 3.55\,\mathrm{dB}$. In other words, parameter estimation is only 1.5 dB short of parameter quantization in terms of spectral distortion. Compared to previous investigations that, typically, report distortions between 6 and 8 dB, an improved speech quality can be expected from the proposed ABWE scheme.

Subjective Speech Quality

Subjectively, it can be stated that the proposed algorithm for artificial bandwidth extension is capable of producing a high band signal with few artifacts only. Overall, the method can regenerate the impression of a real wideband signal. Naturally, when compared with a *coded* version, it can be noticed that the sound is still a little "muffled" and lacks some brilliance which can be attributed to the inevitably limited dynamic range in the estimate of the frequency envelope. However, certain phonemes that are sometimes problematic for ABWE algorithms, especially /*s*/ and /*f*/, seem to be identified correctly in most cases which can be explained by the comparatively high band split frequency of 4 kHz which significantly increases the mutual information measure, see Section 4.3.3.

4.3.6 Application in 3GPP EFR

In an alternative implementation, the TDBWE algorithm with parameter estimation can also be applied to the 3GPP EFR codec which is predominantly used for mobile telephony in the GSM and UMTS networks. Here, compared to the previous description, several modifications have been applied which have been experimentally found to improve the quality of the regenerated extension band signal. This codec is evaluated in Chapter 6 as "Codec under Test B" (CuT-B).

Reduced Parameter Set

In addition to the reduced temporal envelope according to (4.10), also the spectral envelope is represented in a more compact way, i.e., several frequency bands $F(\lambda, m)$ are merged. The reduced parameter set of dimension five is defined as:

$$\mathbf{p}(\lambda) = \left(\bar{T}(\lambda), \bar{F}(\lambda, 0), \bar{F}(\lambda, 1), \bar{F}(\lambda, 2), \bar{F}(\lambda, 3)\right)^{\mathrm{T}} \tag{4.14}$$

with

$$\bar{F}(\lambda, i) \doteq \frac{F(\lambda, 3i) + F(\lambda, 3i+1) + F(\lambda, 3i+2)}{3} - \bar{T}(\lambda). \tag{4.15}$$

In (4.15), the averaged temporal envelope parameter $\bar{T}(\lambda)$ is subtracted from the (averaged) spectral envelope parameters to eliminate redundant gain information.

Additional Feature

For the implementation with the 3GPP EFR as baseband codec, also the *zero crossing rate* (ZCR) is considered as a scalar feature to support the distinction between voiced and unvoiced speech sounds:

$$\mathrm{ZCR} \doteq \frac{1}{2 \cdot (L-1)} \sum_{i=1}^{L-1} \Big| \operatorname{sign}(s_{\mathrm{nb}}(k-1)) - \operatorname{sign}(s_{\mathrm{nb}}(k)) \Big|. \tag{4.16}$$

Moreover, a KLT is applied to decorrelate the feature vector $\mathbf{x}_{\mathrm{f}}(\lambda)$.

Parameter Post-Processing

The simple post-processing of Section 4.3.4 is replaced by a more sophisticated procedure. In particular, the estimated gain parameter $\hat{T}_{\mathrm{MMSE}}$ is post-processed by the following steps:

- Low values of $\hat{T}_{\mathrm{MMSE}}$ are further attenuated to reduce unwanted fluctuations in low-energy segments.
- Strong onsets of $\hat{T}_{\mathrm{MMSE}}$ are attenuated if the baseband signal has a low gain.
- The attenuated value of $\hat{T}_{\mathrm{MMSE}}$ is filtered with an adaptive smoothing filter.
- The resulting value is bounded to the lower end.

In addition to the temporal processing, the estimated *spectral* $\hat{F}_{\mathrm{MMSE}}$ gains are smoothed over frequency with a short gliding average. In effect, this postprocessing is able to remove most transient artifacts in the estimated extension band signal.

4.4 Estimation of G.729.1-SWB Parameters

Another attractive application of artificial bandwidth extension techniques is the extension of wideband speech signals towards the super-wideband bandwidth, a topic which has not been broadly studied so far. Here, an example implementation to estimate the super-wideband parameter set of Section 3.3 is briefly described. This algorithm proposal is evaluated in Chapter 6 as CuT-E.

4.4.1 Reduced Parameter Set

The parameter set which is used for statistical model training and estimation is composed similar to the reduced TDBWE parameter set from the previous section. Concretely, a single temporal gain for each 20 ms frame is used together with a reduced set of spectral subband gains. As in (4.15), groups of three adjacent subbands are merged, resulting in the gains of five broader subbands. Finally, the tonality parameter $\tau(\lambda)$ (Section 3.3.1) is added to the (7-dim.) parameter vector $\mathbf{p}(\lambda)$.

It is worth noting that the binary flags $t(\lambda)$ (transient indicator) and $f(\lambda)$ (replication mode indicator) are not included in the estimation process. Instead, *all* estimated frames are labeled as transient ($t(\lambda) = 1$). Moreover, the spectral replication mode ($f(\lambda) = 1$) is *always* used to regenerate the spectral details in the extension band which is fully sufficient for speech signals. As a consequence, the harmonic pitch grid and offset parameters ($p(\lambda)$ and $p_{\text{offset}}(\lambda)$) can be disregarded entirely, cf. Section 3.3.3.

4.4.2 Wideband Features

The wideband feature vector $\mathbf{x}_{\text{f}}(\lambda)$ comprises three principal subvectors:

- The first subvector is chosen similar to the narrowband features from Section 4.3.2, i.e., the *line spectrum pairs* (LSPs) of the G.729 *narrowband* core codec are directly used as features. Also, the zero crossing rate (4.16) of the *narrowband* signal is added to the feature vector. The temporal envelope of the narrowband signal is represented with a single gain.

- The second part of the feature vector is a direct copy of the TDBWE parameter set (Section 3.2.1) which describes the temporal and spectral structure of the *first* extension band in the range from 4 to 7 kHz.

- Finally, also the gains of the *subsequent* 20 ms frame, both for the narrowband (0 – 4 kHz) and for the 4 – 7 kHz extension band, are added as features. This is possible without any additional delay because the G.729.1-SWB codec operates in the MDCT domain. The inverse transform requires one frame of additional delay because of the final overlap-add step (2.28). The inclusion of "future" gains in the feature vector leads to a more consistent temporal evolution of the estimated parameters because of the additional look-ahead.

As in the TDBWE case, a KLT is applied to decorrelate the vector $\mathbf{x}_{\text{f}}(\lambda)$.

4.4.3 Parameter Post-Processing

The parameter post-processing closely follows the procedure as described in Section 4.3.6. Concretely, the estimated temporal gain is post-processed by adaptive smoothing, an attenuation of lower energies and limiting. Only the attenuation of strong onsets is not needed because the estimation of super-wideband parameters is more reliable than the estimation of wideband parameters.

The estimated spectral envelope parameters are also smoothed with a short gliding average filter. Since the estimated signal exhibits a slightly noisy characteristic, the estimated tonality parameter $\hat{\tau}_{\text{MMSE}}(\lambda)$ is weighted towards a more tonal signal regeneration.

4.5 Discussion

In this chapter, the receiver based estimation of bandwidth extension parameters *within* the framework of a standardized codec has been proposed. Therefore, the signal processing algorithms for bandwidth extension that have been described in Chapter 3 are reused. The considered setup is interesting in a number of potential applications, for example:

- In a heterogeneous conference scenario which is based on a hierarchical audio codec such as ITU-T G.729.1, there might be users who perceive a *mix* of narrow- and wideband speech which impairs the conversational quality. Here, artificial bandwidth extension, i.e., estimation of the missing extension band parameters, can provide a constant bandwidth.

- Bit rate switchings due to network congestion may occur between different bit rates of a hierarchical codec, e.g., between 12 and 14 kbit/s in G.729.1. Normally, this causes severe artifacts as the audio bandwidth is significantly narrowed during a short period. Appropriately inserting an estimated high band signal during periods of network congestion can virtually remove such artifacts. This approach, initially proposed in [Geiser et al. 2006] for a G.729.1 candidate codec, is also used in [Ramabadran & Jasiuk 2008, Laaksonen et al. 2010] for the G.729.1 super-wideband extension.

Compared to a dedicated ABWE extension band synthesis algorithm, no detrimental effects on speech quality could be observed when reusing the *standardized* method for extension band parametrization and synthesis. It should also be mentioned that the reuse of a standardized extension band synthesis algorithm is advantageous in terms of computational complexity because the required amount of *additional* complexity is entirely determined by the parameter estimation procedure.

In the example implementation of Section 4.3, the TDBWE parameters of the G.729.1 codec (see Section 3.2.1) are estimated based on the narrowband bitstream layers of this codec. The respective measurements of mutual information and "high band certainty" (Section 4.3.3) have shown that the TDBWE parameters are actually suitable for a concise estimation. It was found that a satisfactory wideband speech quality can be obtained. Especially the particularly critical fricative sounds can be identified correctly in most cases.

A basic problem of ABWE algorithms is that only a few look-ahead samples are allowed for an application in real-time, bidirectional telephony. Where, for instance, a speech recognition system may segment the input signal on the phoneme, word, or even sentence level, the ABWE estimator has to decide "on the spot" how much energy to put in the extension band. It is an inherent problem of current ABWE estimation algorithms that the resulting (unavoidable)

estimation errors are not taken into account in the estimation of the extension band parameters in *future* signal frames. For example, an *under-estimation* of the extension band gain in frame λ might not be perceived as an artifact by itself, but with the placement of the *correct* amount of energy in the *following* frame $(\lambda + 1)$, an artifact (late onset) could be produced. Here, it is proposed to mitigate such detrimental effects with a parameter post processing procedure, which is particularly important for the extension band gain parameter. Moreover, for an implementation of an ABWE scheme in the MDCT domain (Section 4.4), "future" baseband features are available. Their exploitation also contributes to a more consistent evolution of the estimated high band parameter set.

As another new aspect, the estimation of a *super-wideband* parameter set from an available *wideband* signal has been proposed. The example implementation is based on the G.729.1-SWB candidate (Section 3.3). The extension of wideband speech towards the super-wideband bandwidth in fact yields better and much more consistent estimation results than the typical narrowband to wideband extension. This observation is substantiated by the listening test results in Chapter 6.

4.6 Comparison with Other Approaches

To complete this chapter, the most prominent techniques that have been investigated and proposed for artificial bandwidth extension of speech signals shall be summarized here. More resources and reviews of the literature published on the topic can be found in [Jax 2002, Larsen & Aarts 2004, Iser et al. 2008, Pulakka & Alku 2011]. Note that the vast majority of ABWE proposals, are in fact limited to *speech* signals. A blind bandwidth extension of *music signals* has not been broadly investigated. Nevertheless, a few proposals exist, e.g., [Larsen et al. 2002, Liu et al. 2003].

Signal Processing Techniques

In contrast to the large number of different algorithms for parametric (speech) bandwidth extension in embedded coding frameworks (as presented in Section 3.4), the *signal processing techniques* to synthesize the extension band components are usually less varied in ABWE proposals. Often, a simple autoregressive synthesis of the spectral envelope is applied based on an excitation signal that is a spectral mirror image of the baseband signal. Nevertheless, there are also some approaches that use other synthesis techniques such as a filterbank equalizer with non-uniform frequency resolution to shape the spectral envelope of the signal.

Features

The general approach for the estimation of the extension band parameters is very similar in virtually all proposals, i.e., a number of relevant *features* is extracted

from the baseband signal and mapped to the missing extension band parameters using pre-trained (statistical) mapping functions. Typical baseband features are the spectral envelope, e.g., in the form of autoregressive (AR) coefficients, Mel-Frequency Cepstral Coefficients (MFCCs) or subband energies, but also features describing other signal characteristics such as the spectral fine structure (zero crossing rate, gradient index, etc.). An analysis of a number of common baseband features and of their impact on the estimation performance has been published in [Jax & Vary 2004]. In particular, the use of MFCC features and of the derived so called "delta features" is investigated in [Nour-Eldin et al. 2006, Nour-Eldin & Kabal 2009, Nour-Eldin & Kabal 2011].

Estimation Methods

The main point for differentiation between different ABWE algorithms is the employed method for parameter estimation. Many estimation methods have been investigated and proposed in the literature. The more prominent proposals shall be listed here.

The mapping of the entries of a narrowband codebook to a wideband *shadow codebook* has been proposed in [Carl & Heute 1994]. Some other approaches which are based on the idea of *codebook mapping* are [Enbom & Kleijn 1999, Unno & McCree 2005] and [Kornagel 2006].

The second class of mapping functions employs a *(piecewise) linear mapping* of the feature vectors to the extension band parameter vectors, see [Nakatoh et al. 1997, Chennoukh et al. 2001]. Another mapping based approach, termed "feature mapping," has been proposed in [Gustafsson et al. 2001]. As an alternative, *artificial neural networks* are used to map the features to the parameters in [Kontio et al. 2007, Pham et al. 2010, Pulakka & Alku 2011].

In contrast to the mapping based approaches, *statistical* estimation schemes rely on the modeling of the underlying probability densities. Maximum-a-posteriori (MAP) or Minimum-mean-square-error (MMSE) estimators can then be explicitly formulated. The use of *Gaussian mixture models (GMMs)* has been proposed in [Park & Kim 2000]. GMM-based techniques are also used in various other publications, e.g., [Nour-Eldin et al. 2006, Kim et al. 2008, Nour-Eldin & Kabal 2009, Pulakka et al. 2011]. As a modification of the classical GMM based estimator, an asymmetric cost function can been adopted [Nilsson & Kleijn 2001], avoiding an overestimation of the extension band energy.

With statistical estimation methods, also the temporal evolution of the extension band parameter set can be explicitly modeled by using *Hidden Markov Models (HMMs)*. The HMM estimation method, as described in Section 4.2, has been initially proposed for ABWE in [Jax & Vary 2000], see also [Jax 2002, Jax & Vary 2003]. The *language dependency* of the HMM estimation scheme has been investigated in [Bauer & Fingscheidt 2008], revealing that a *multi-lingual training* is feasible whereafter even languages that have not been included in the training

process can be successfully processed. Other HMM based estimators that try to model larger blocks (e.g. phones) have been described in [Yao & Chan 2005] and [Yağlı & Erzin 2011]. A comparison between GMM and HMM based estimation has been conducted in [Song & Martynovich 2009]. As an alternative to Hidden Markov modeling, the temporal evolution of the estimated parameters can also be accounted for by *Kalman filtering*, cf. [Yao & Chan 2006].

Most publications mentioned above only deal with the bandwidth extension towards higher frequencies (usually 3.4 – 7 kHz). However, there are also some proposals that try to regenerate the missing *low frequencies* of telephone speech, i.e., the frequency range from 50 – 300 Hz, see [Miet et al. 2000, Valin & Lefebvre 2000, Kornagel 2003, Park et al. 2004, Thomas et al. 2010].

Applications

Meanwhile, a number of real-world application of ABWE techniques have been reported. First, the 3GPP AMR-WB codec [ETSI 2001*b*, Bessette et al. 2002] artificially regenerates the (relatively narrow) extension band from 6.4 – 7 kHz by shaping synthetic noise. A full solution for narrow-to-wideband bandwidth extension in mobile phones is presented in [Pulakka et al. 2008, Laaksonen et al. 2009]. The application of ABWE in an automotive environment is discussed in [Bauer et al. 2010].

Another application of artificial bandwidth extension emerges in the context of embedded audio codecs (Chapter 3). If the wideband bitstream layers have to be truncated from the bitstream, e.g., due to intermediate network congestion, a temporary drop of the reproduced audio frequency can be avoided by inserting an artificially regenerated signal into the extension band. This idea, as discussed in Section 4.5, has been implemented in [Geiser et al. 2006] and [Ramabadran & Jasiuk 2008, Laaksonen et al. 2010].

Surprisingly, artificial bandwidth extension techniques also appeared in a number of other applications, for example in wideband noise suppression which can be supported by estimated higher audio frequencies, especially in low SNR conditions [Esch et al. 2010]. An estimate for the extension band parameters can also help to achieve a low bit rate in predictive parameter quantization schemes, cf. [Ehara et al. 2005, Agiomyrgiannakis & Stylianou 2007]. The application of ABWE to binaural signals has been discussed in [Laaksonen & Virolainen 2009].

Chapter 5

Steganographic Parameter Transmission

Unfortunately, the purely receiver-based algorithms for artificial bandwidth extension, as described in the previous chapter, cannot deliver a sufficiently stable wideband or super-wideband speech quality under all circumstances. Also the more robust solution of Chapter 3, i.e., embedded coding with quantized bandwidth extension parameters, cannot reliably improve the situation. Here, even if both end-user terminals are suitably equipped, the legacy telephone network (or a legacy section in the transmission chain) will discard any enhancement bits and therefore effectively preclude the high quality audio reproduction at the receiver.

A new solution to resolve this dilemma is discussed here. Based on the parametric bandwidth extension techniques of Chapters 2 and 3, it is proposed to communicate information about the missing audio frequencies over a *steganographic* channel, i.e., the related bits are *hidden* within the narrowband speech signal or within the legacy bitstream using *data hiding* or *watermarking* techniques. The bitstream format of the legacy codec is not altered and the bit rate is not increased. The modified bitstream can still be decoded by any standard narrowband decoder whereby, naturally, only a very limited loss in terms of narrowband speech quality can be accepted. However, an enhanced decoder which is aware of the hidden information can produce a wideband speech signal of much higher quality.

In this chapter, after a review of data hiding fundamentals (Section 5.1), the combination (and interaction) of data hiding *and* source coding within a single transmission system is investigated using three conceptually different approaches (Section 5.2). The application to speech and audio transmission systems is discussed in Section 5.3. Thereby, the principle of "joint source coding and data hiding" is shown to be particularly relevant for the present application and especially for state-of-the-art ACELP speech codecs (Section 5.4). The devised methods of hidden data transmission are exemplarily applied to two standardized and widely deployed narrowband speech codecs (Section 5.5). As the final step in Section 5.7, a bandwidth extension algorithm (cf. Section 3.2) is added to a narrowband codec and the respective parameters are transmitted over the steganographic channel. The resulting transmission system is *backwards compatible* w.r.t. legacy narrowband terminals and the network itself. As an application example, the transmission of hidden information over a standard GSM cellular network has been simulated.

5.1 Data Hiding

In a digital communication system, techniques for "data hiding" or "digital watermarking" allow to establish a virtual communication channel that is hidden within the transmitted "host signal" without increasing the bit rate. In practice, the host signal usually represents multimedia data, i.e., audio, image, or video signals. The data can then be hidden directly in the *signal samples*, in the *transform coefficients* or within a *parametric* description of the multimedia content. An excellent overview of data hiding theory and the related techniques is provided in [Moulin & Kötter 2005] and [Cox et al. 2008]. Here, a few essential ideas and facts shall be summarized.

5.1.1 Fundamentals

A generic model for a data hiding system is shown in Figure 5.1. The general task of data hiding is to embed a message m taken from a set of possible messages $\mathbb{M} \doteq \{0, 1, \ldots, M-1\}$ into a host signal (vector) $\mathbf{x} \in \mathbb{R}^n$ by applying an *embedding function* $\tilde{\mathbf{x}} = f(\mathbf{x}, m, \mathbf{k})$ which may also depend on a key vector $\mathbf{k}$ if cryptographic security is desired. The modified signal $\tilde{\mathbf{x}}$ has to be (in some sense) similar to the original host signal $\mathbf{x}$ while the message m must remain recoverable from $\tilde{\mathbf{x}}$ or even from a disturbed version $\mathbf{y} = \tilde{\mathbf{x}} + \mathbf{n}$ of the signal with the (effective) additive noise term $\mathbf{n}$. The decoded hidden message is denoted by $\hat{m} \in \mathbb{M}$.

There are two sources of distortion that play important roles in data hiding. The first one is the so called *embedding distortion* $d(\mathbf{x}, \tilde{\mathbf{x}})$ that is introduced by the embedding function itself. The second source of distortion is the *channel distortion* $d(\tilde{\mathbf{x}}, \mathbf{y})$. Thereby, a particular distortion measure $d(\cdot, \cdot)$ must be chosen to weight the introduced error appropriately. For easier theoretical analysis and in many practical applications, the squared Euclidean distance $d(\mathbf{x}_1, \mathbf{x}_2) = ||\mathbf{x}_2 - \mathbf{x}_1||^2$ is a common choice. Based on the Euclidean error measure, two quantities can be defined to characterize a data hiding system:

1. The *host-to-watermark-ratio* $\text{HWR} = \frac{||\mathbf{x}||^2}{d(\mathbf{x}, \tilde{\mathbf{x}})}$ quantifying the "embedding strength" and
2. the *watermark-to-noise-ratio* $\text{WNR} = \frac{d(\mathbf{x}, \tilde{\mathbf{x}})}{d(\tilde{\mathbf{x}}, \mathbf{y})}$ to characterize the robustness against channel noise.

A *good* data hiding scheme has to be designed such that

- the hidden data can be detected and/or extracted reliably at the receiving end (possibly even after deliberate "attacks" or attempts to destroy the watermark), i.e., the WNR must be sufficient,
- the modified host signal $\tilde{\mathbf{x}}$ is not (or hardly) subjectively distinguishable from the original signal $\mathbf{x}$ (which corresponds to high HWR values),
- and a minimum hidden data rate ($R_{\text{DH}} \doteq \text{ld}\, M$ bit/vector) is guaranteed.

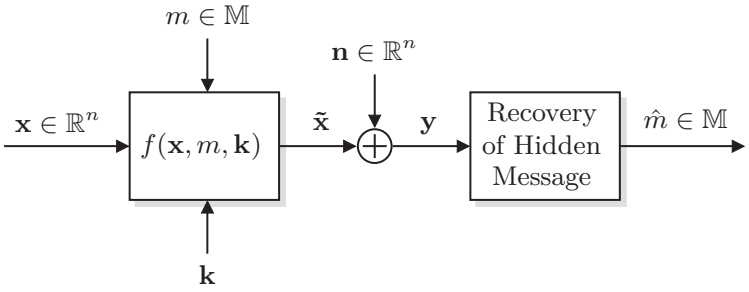


Figure 5.1: Generic model of a data hiding system.

Hence, the challenge in designing good data hiding schemes is to find an adequate compromise between the (contradicting) requirements of high data rate, sufficient robustness to noise, and low amount of distortion introduced into the host signal.

A widespread application of data hiding is the indication of the host signal's origin, e.g., for authentication purposes, copyright protection or digital rights management. In contrast, this thesis aims to transmit *auxiliary data* which is then used to *enhance the host signal*. In this case, the robustness to deliberate attacks might be less relevant, but other transmission characteristics are more important such as a higher hidden data rate R_{DH}, the need for a *constant* (minimum) rate, and robustness to transmission errors. Therefore, with view to the desired application, the scenario of a deliberate attack as well as the cryptographic security of the transmitted message are disregarded in the following.

5.1.2 Data Hiding Based on the Principle of Binning

The data hiding methods to be employed in this thesis are based on the principle of "binning" [Cox et al. 2008, Zamir et al. 2002] which can be interpreted as (vector) quantization, e.g., [Gray & Neuhoff 1998], with prior codebook selection.

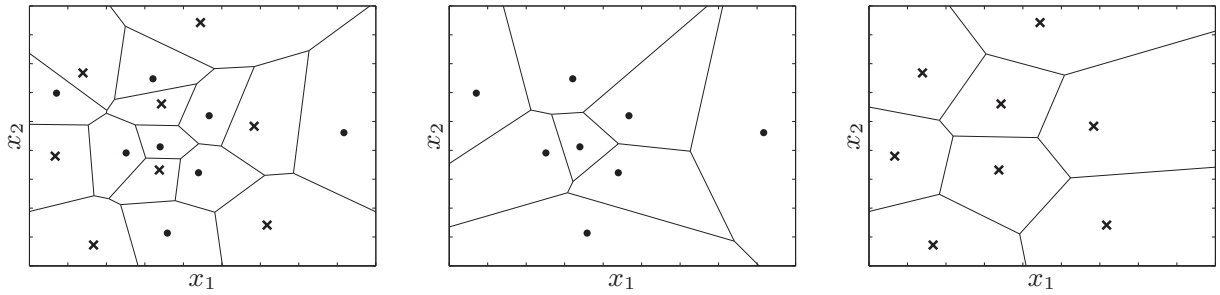
Vector quantization can, in general, be described as a mapping of an input vector $\mathbf{x} \in \mathbb{R}^n$ to a representative $\hat{\mathbf{x}}$ from a quantizer codebook $\mathcal{C} \doteq \{\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1, \ldots\}$ such that the quantization distortion $d(\mathbf{x}, \hat{\mathbf{x}})$ is minimized, i.e.,

$$\hat{\mathbf{x}} = \arg\min_{\hat{\mathbf{x}}' \in \mathcal{C}} d(\mathbf{x}, \hat{\mathbf{x}}'). \tag{5.1}$$

An exemplary codebook for vector quantization of 2-dimensional input vectors $\mathbf{x} = (x_1, x_2)^{\mathrm{T}}$ is shown in Figure 5.2(a). The codebook entries $\hat{\mathbf{x}}_0, \hat{\mathbf{x}}_1, \ldots$, i.e., the *centroids* of the quantization cells are marked (dots or crosses).

To hide R_{DH} bits of information in the host vector $\mathbf{x} \in \mathbb{R}^n$, in total $M = 2^{R_{\mathrm{DH}}}$ *bins*, i.e., *disjoint codebooks* $\mathcal{C}_m$ with $m \in \mathbb{M}$ are required. To be disjoint, the M codebooks $\mathcal{C}_m$ must fulfill

$$\mathcal{C}_m \cap \mathcal{C}_{m'} = \emptyset \quad \text{for} \quad m \neq m' \quad \text{with} \quad m' \in \mathbb{M}. \tag{5.2}$$



(a) Codebook $\mathcal{C}$ with sub-codebook labels: $\bullet$, $\times$.

(b) Sub-codebook $\mathcal{C}_\bullet$ for data hiding.

(c) Sub-codebook $\mathcal{C}_\times$ for data hiding.

Figure 5.2: Example for data hiding based on binning.

As one possible approach to establish the individual codebooks $\mathcal{C}_m$, *codebook partitioning* can be used. In this case, the initial (extensive) codebook $\mathcal{C}$ is partitioned into M disjoint *sub-codebooks* $\mathcal{C}_m$. An example partitioning for two sub-codebooks $\mathcal{C}_\bullet$ and $\mathcal{C}_\times$ ($M = 2$) is shown in Figure 5.2, allowing a hidden data rate of $R_{\mathrm{DH}} = 1\,\mathrm{bit/vector}$. In general, when the codebook partitioning approach is used, the following property must hold in addition to (5.2):

$$\bigcup_{m \in \mathbb{M}} \mathcal{C}_m \subseteq \mathcal{C}. \tag{5.3}$$

A particular message $m_0 \in \mathbb{M}$ can then be hidden in the vector $\mathbf{x}$ by selecting the appropriate codebook for quantization, i.e., the codebook search procedure *with information-embedding* is, in analogy to (5.1), defined as

$$\tilde{\mathbf{x}} = \arg \min_{\hat{\mathbf{x}} \in \mathcal{C}_{m_0}} d(\mathbf{x}, \hat{\mathbf{x}}). \tag{5.4}$$

To recover m_0 at the receiver side, the decoder only needs to identify the specific codebook $\mathcal{C}_{m_0}$ that has been used to produce $\tilde{\mathbf{x}}$, i.e., the hidden message m_0 is given by

$$m_0 = \{m \,:\, \tilde{\mathbf{x}} \cap \mathcal{C}_m = \tilde{\mathbf{x}}\}. \tag{5.5}$$

If the transmission channel is noisy, i.e., an "attack" noise $\mathbf{n}$ is involved, a dedicated decoder algorithm is required. In the simplest case, nearest neighbor decoding of $\mathbf{y} = \tilde{\mathbf{x}} + \mathbf{n}$ over all codebooks $\mathcal{C}_m$ can be used to recover the hidden message

$$\hat{m}_0 = \arg \min_{m \in \mathbb{M}} \min_{\hat{\mathbf{x}} \in \mathcal{C}_m} d(\mathbf{y}, \hat{\mathbf{x}}), \tag{5.6}$$

whereby decoding errors ($\hat{m}_0 \neq m_0$) may occur depending on the noise $\mathbf{n}$. Equation (5.6) corresponds to the *maximum likelihood estimate* of the hidden message for a well-behaved distribution of the noise $\mathbf{n}$.

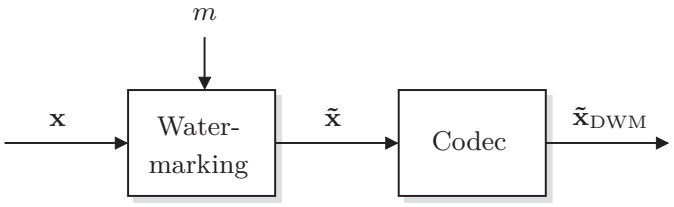
5.1.3 Properties of Good Data Hiding Codes

Many state of the art methods for data hiding can be interpreted as variants of the above described binning principle. Binning has become relatively popular since the related code designs are known to be asymptotically optimal in achieving the *Shannon capacity* $C_0 = \frac{1}{2}\mathrm{ld}\,(1 + \mathrm{WNR})$ of the (ideal) steganographic communication channel [Costa 1983, Moulin & Kötter 2005]. It is remarkable that, to achieve the channel capacity, the host signal $\mathbf{x}$ can remain unknown to the decoder. Hence, the impact of interference at the transmitter (i.e., of the host signal) can be completely eliminated by an appropriate choice of the data hiding code. Therefore, the capacity C_0 does not depend on the HWR and is only determined by the WNR, see also [Costa 1983, Cox et al. 1999, Erez & Zamir 2004, Moulin & Kötter 2005].

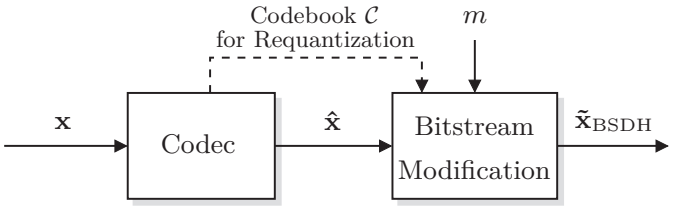
However, there is still the question how to *jointly* design the codebooks $\mathcal{C}_m$ and the union-codebook $\mathcal{C}$ in Equation (5.3). Since, according to (5.4), the embedding distortion, and therefore the HWR, is entirely determined by the properties of the sub-codebooks $\mathcal{C}_m$, these must, individually, form good *quantization codebooks* that guarantee a low quantization error, e.g., [Gray & Neuhoff 1998]. In the case of a uniformly distributed source vector $\mathbf{x}$ of dimension n, this can be fulfilled by quantization cells with a shape that is as close as possible to the n-dimensional sphere. The union-codebook $\mathcal{C}$ must, on the other hand, fulfill a different design constraint. The objective here is to cope with the (effective) channel noise $\mathbf{n}$. This can, for instance, be achieved by maximizing the minimum pairwise distance between the individual codebook entries. Eventually, other system requirements (such as backwards compatibility) have to be regarded.

Naturally, the design constraints for the sub-codebooks $\mathcal{C}_m$ and the union-codebook $\mathcal{C}$ do not necessarily coincide and an adequate tradeoff has to be found for a given application. In theory, code designs based on *random* binning principles, i.e., random assignment of sub-codebook labels "m" to the entries of the (also random) codebook $\mathcal{C}$, have been shown to be capacity achieving [Costa 1983]. However, such codes are usually not usable in practical applications because of storage and complexity constraints.

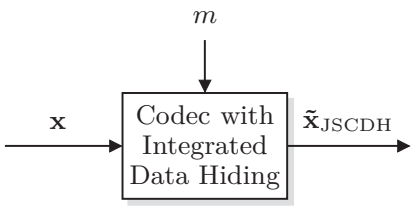
To overcome the practical difficulties of random codes, many practically usable data hiding codes are based on *structured* approaches involving algebraic quantization methods. A popular example is the method of "quantization index modulation" (QIM) as introduced by [Chen & Wornell 2001]. A variant thereof, using scalar quantization, is known as the "Scalar Costa Scheme" [Eggers et al. 2003]. Other authors have proposed QIM schemes based on multi-dimensional *lattice quantization*, e.g., [Zhang & Boston 2003, Fischer & Bäuml 2004, Geiser et al. 2005].



(a) Digital Watermarking (DWM)



(b) Bitstream Data Hiding (BSDH)
Variant 1: LSB replacement
Variant 2: Requantization



(c) Joint Source Coding and Data Hiding (JSCDH)

Figure 5.3: Concepts for data hiding in combination with source coding.

5.2 Data Hiding and Source Coding

If data hiding is to be combined with a source coding system, it is a strict requirement that the codewords to be transmitted over the channel are taken from a *given* source codebook $\mathcal{C}$ that is associated with the respective transmission standard. Then, three conceptually different approaches, as illustrated in Figure 5.3, are conceivable. These systems are discussed in the following.

1. A data hiding algorithm can be directly applied to the (transformed) host signal vector $\mathbf{x}$, see Figure 5.3(a). This approach is referred to as *digital watermarking* (DWM). The watermarked signal $\tilde{\mathbf{x}}$ is encoded *after* data embedding and then transmitted over the communication system. Particularly, as the watermarking algorithm is assumed to be unaware of the subsequent source en- and decoding, any coding noise directly impacts the watermark detection and decoding. Hence, coding noise must be regarded as a malicious attack signal and watermark decoding errors may occur. Moreover, it must be ensured by the concrete system design that the *overall* SNR $d(\mathbf{x}, \tilde{\mathbf{x}}_{\mathrm{DWM}})$ is not increased exceedingly.

2. In contrast to the "classical" DWM approaches, steganographic data can alternatively be embedded into a *compressed* or *encoded* representation of the host signal. This method is sometimes called "bitstream watermarking" or "compressed domain watermarking." Naturally, it is only applicable if the considered transmission system implements signal compression, for

instance a speech codec. The data embedding is then performed directly on the content of the bitstream, in the simplest case by *overwriting least significant bits* (LSBs), or, in a more sophisticated manner, by *requantizing the previous reconstruction vector* with the respective sub-codebook $\mathcal{C}_m$. The respective system setup is shown in Figure 5.3(b). Although this approach—here referred to as "bitstream data hiding" (BSDH)—is obviously immune to coding noise, it is still suboptimal. Especially if low-rate source encoding is used, LSBs might still be too significant to be altered without causing severe quality degradation.

3. The third approach in Figure 5.3(c) exploits the fact that data hiding and source encoding are often co-located. This facilitates a *joint* implementation and optimization of source encoding *and* data hiding. Consequently, this method is referred to as "joint source coding and data hiding" (JSCDH). This system has several advantages as shown in the following.

Obviously, the choice for the M data hiding codebooks $\mathcal{C}_m$ (see Section 5.1.2) for the latter two systems, i.e., for BSDH and JSCDH, is restricted to subsets of the given source codebook $\mathcal{C}$ if full bitstream compatibility with the given transmission system must be maintained. In contrast, the data hiding codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ for the DWM setup may be chosen arbitrarily, but they may *not* exploit explicit knowledge about the source codebook $\mathcal{C}$. Otherwise, the watermarking algorithm in Figure 5.3(a) could simply produce vectors $\tilde{\mathbf{x}}$ that are directly taken from $\mathcal{C}$. Such a system would, in effect, be identical to the JSCDH approach.

The effect of DWM, BSDH, and JSCDH is qualitatively illustrated in Figure 5.4 based on simple scalar quantization with natural binary index assignment. It is assumed that the scalar host signal, i.e., the parameter or sample value $x \in \mathbb{R}$, is available with a resolution that is better than the resolution of the quantizers. For the transmission of $\mathrm{ld}\, M = 1$ hidden bit, the data hiding codebooks are established by dividing the set of quantizer reproduction levels into $M = 2$ subsets with even or odd indices (binning). The transmitted information is then represented by the choice of the subset. As the data hiding codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ for DWM must not coincide with the source codebook $\mathcal{C}$, an (arbitrary) offset between the respective centroids has been assumed in the DWM example.

In Figure 5.4, DWM and the subsequent source coding result in the index value 101. The incurred distortion is $d(x, \tilde{x}_{\mathrm{DWM}})$. Instead, the BSDH approach changes the *initial* quantizer index 100 to 101 by overwriting the LSB, leading to a distortion of $d(x, \tilde{x}_{\mathrm{BSDH}})$. In contrast, the JSCDH result 011 is immediately obtained with a *single* quantization step. Of all three approaches, conforming to intuition, JSCDH yields the lowest embedding distortion $d(x, \tilde{x}_{\mathrm{JSCDH}})$. This is also supported by the theoretical results of [Cohen et al. 2006].

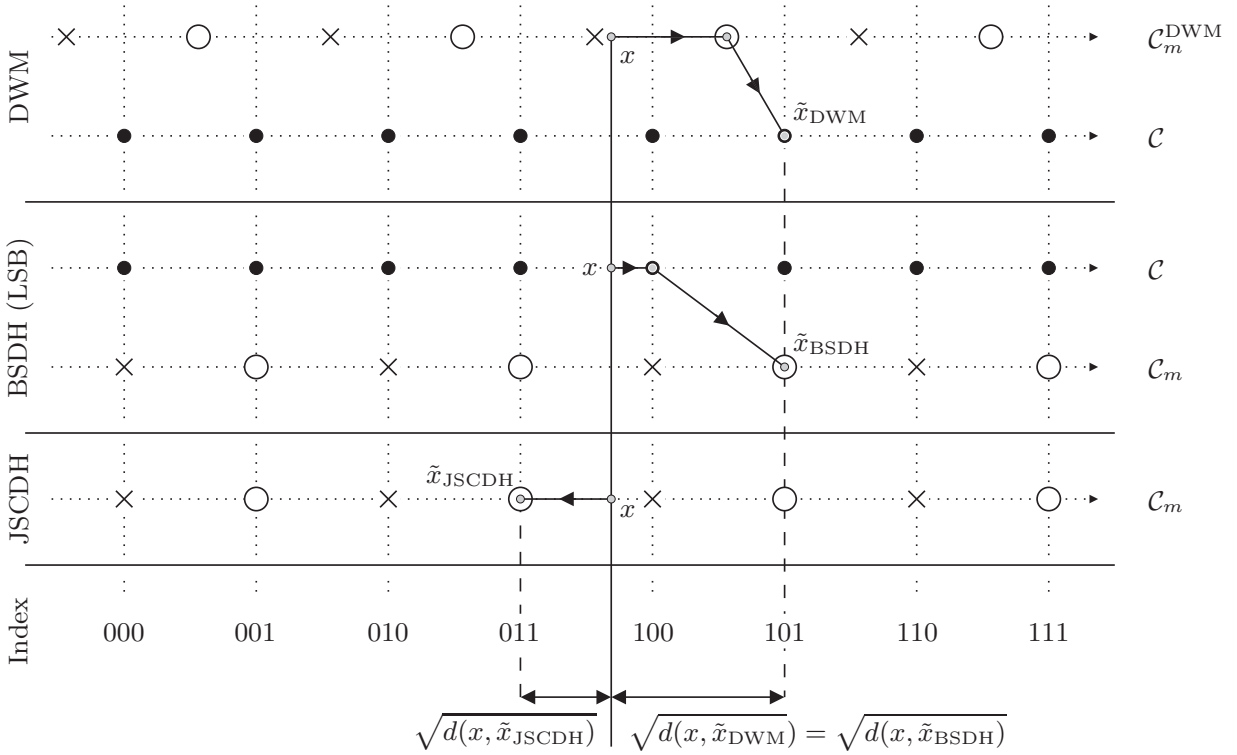
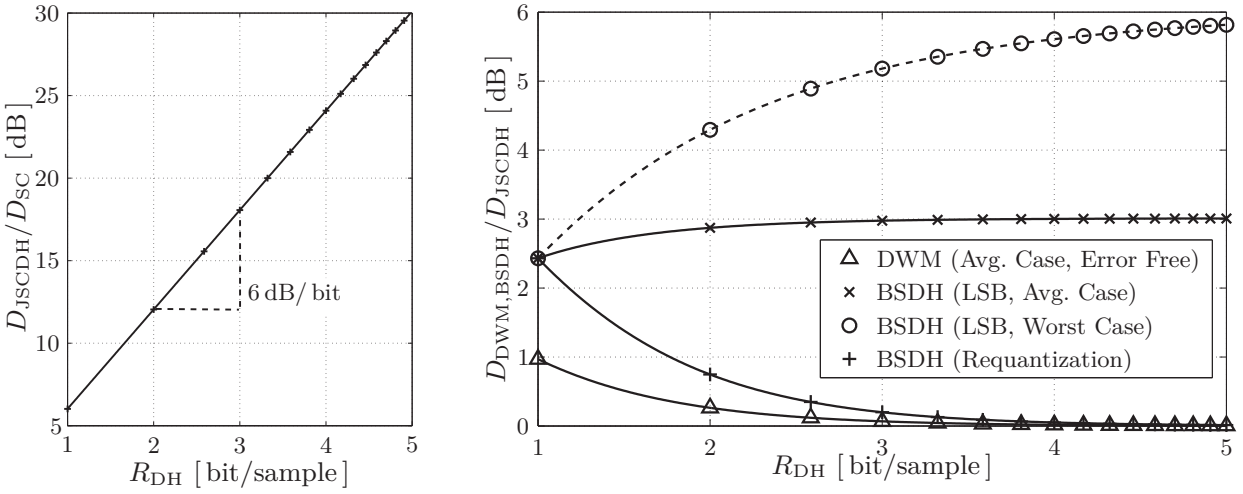


Figure 5.4: Qualitative comparison of DWM, BSDH, and JSCDH with 3-bit scalar quantization. One hidden bit $m = 1$ is transmitted. Definition of the sub-codebooks: $\mathcal{C}_0 = \mathcal{C}_\times$, $\mathcal{C}_1 = \mathcal{C}_\circ$. DWM and BSDH lead to equal distortion in the shown case.

Figure 5.5 provides a more quantitative comparison of the considered scenarios in terms of the mean squared error values $D_{(\cdot)} \doteq \mathrm{E}\left\{d(x, \tilde{x}_{(\cdot)})\right\}$ based on scalar quantization. The analytical derivation of these results is deferred to Appendix B. Figure 5.5(a) illustrates the distortion penalty that is incurred by JSCDH compared to mere source coding with the source codebook $\mathcal{C}$. The *additional* penalty of BSDH and DWM compared to JSCDH is quantified in Figure 5.5(b). It is important to note that, to allow a fair comparison with BSDH and JSCDH, the DWM codebooks have been chosen such that the probability of a decoding error $P(m \neq \hat{m})$ is *exactly* zero[1] while the embedding distortion is minimized. An alternative design choice for DWM is to enforce the same embedding distortion as for JSCDH and to accept a certain non-zero error probability. For BSDH, both the requantization approach (which is feasible if the source codebook is available in the data hiding unit) and the LSB replacement method are analyzed. The "worst case" in LSB replacement occurs if the original LSBs are zero, see Appendix B. The BSDH system (even with requantization) is inferior to DWM in the considered scenario because the latter system has access to the *original* host signal x.

[1]This is also the case for the DWM system in Figure 5.4, because the stepsize of the codebook $\mathcal{C}$ is not larger than the stepsize of the union of $\mathcal{C}_\times^{\mathrm{DWM}}$ and $\mathcal{C}_\circ^{\mathrm{DWM}}$. Requantizing $\tilde{x}_{\mathrm{DWM}}$ with the codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ will *always* yield a correctly decoded message $\hat{m} = m$.



(a) (b)

Figure 5.5: Embedding distortion of DWM, BSDH, and JSCDH for scalar data hiding and source coding. Gaussian source ($\sigma_x^2 \gg \sigma_{x-\tilde{x}}^2$).
(a) JSCDH penalty compared to pure source coding
(b) DWM and BSDH penalty compared to JSCDH

As a conclusion, there is a consistent advantage of JSCDH over DWM and BSDH in terms of embedding distortion, in particular for low hidden bit rates R_{DH}. Although, in the idealized scenario of Figure 5.5(b), this advantage might appear relatively small in some cases, BSDH as well as DWM are expected to perform much worse in *practical* environments. Practical speech and audio coders are much more sophisticated than simple scalar quantizers and the respective LSBs (in the sense of "least *sensitive* bit") are often too important to be blindly replaced by a steganographic message. The application of DWM is also difficult because of the strong compression, i.e., high quantization noise. These points are substantiated in the following sections.

JSCDH also has a considerably lower complexity than the other systems because only a single quantization operation has to be carried out per input vector $\mathbf{x}$. The number of considered codebook entries is only $|\mathcal{C}_m|$, i.e., the complexity is even lower than for mere source coding ($|\mathcal{C}|$). Instead, for DWM and for BSDH with requantization, $|\mathcal{C}_m| + |\mathcal{C}|$ codebook entries must be considered in the two subsequent quantization steps. In contrast, BSDH with substitution of the quantization index LSB, being the most inefficient choice from the distortion perspective, is less complex than BSDH with requantization because the source coder still examines its $|\mathcal{C}|$ codebook entries while the complexity of the LSB substitution itself is negligible. Moreover, several other practical considerations may justify the preference of a JSCDH solution in favor of the DWM/BSDH approaches within speech and audio communication systems. This is discussed in Section 5.3.4.

5.3 Data Hiding in Speech and Audio Communication

For an application of data hiding techniques in digital speech and audio communication systems, several other aspects besides mere embedding distortion need to be taken into account. On the one hand, instead of a simple distortion measure such as the HWR, perceptual criteria can be applied for speech and audio host signals. In particular, auditory masking effects can be exploited to keep the embedding distortion below the perception threshold. Therefore, a lower HWR becomes acceptable which, in turn, increases the WNR. On the other hand, also practical constraints of the specific transmission system must be taken into account. For instance, certain applications (such as, e.g., bandwidth extension) impose a delay constraint and, consequently, information embedding has to be performed *per frame*, i.e., any additional algorithmic delay has to be avoided.

Several variants and realizations of DWM, BSDH, and JSCDH for speech and audio communication systems that have been described in the literature are reviewed in the following. Concrete performance comparisons with the methods that are devised later on in this chapter are given in Section 5.5.3.

5.3.1 Digital Watermarking (DWM)

Mostly, data hiding for speech signals is performed directly on the PCM samples. Common algorithms for speech and audio DWM are "spread spectrum watermarking" [Cheng & Sorensen 2001] and quantization based techniques such as "quantization index modulation" [Chen & Wornell 2001] or the "Scalar Costa Scheme" [Eggers et al. 2003]. These algorithms may alternatively be applied in a transformed domain. In this case, also an inverse transform needs to be implemented to reconstruct the watermarked speech signal from the modified transform coefficients. Such transformations often aim at reduced audibility of the embedded watermark, therefore exploiting auditory masking effects.

Yet, for rather severe transmission conditions such as CELP coding of the marked speech signal with typical source coding bit rates of $4 - 16\,\mathrm{kbit/s}$, it turns out that the hidden message transmission is not very reliable if the watermark is supposed to be imperceptible and hidden data rates of several $100\,\mathrm{bit/s}$ are required. The reason for this is that the CELP transcoding process itself disturbs the embedded message which becomes obvious by recognizing that watermarks, to be imperceptible, are usually embedded into *less relevant components* of the speech signal. Unfortunately, such speech components are also very likely to be coarsely quantized or omitted by the speech coder. For instance, classical spread spectrum watermarks [Cheng & Sorensen 2001] are not even feasible for *very low* hidden bit rates as shown in [Celik et al. 2005].

For a *robust* transmission of hidden data over low-rate CELP codecs, it is desirable to modify *perceptually important speech features* instead, for example the spectral envelope or the pitch structure in voiced speech segments. Only these features are encoded with sufficient accuracy so that a reliable hidden data trans-

mission can be achieved. On the contrary, also here only very low hidden data rates can be achieved if the speech quality shall be maintained. So, obviously, an adequate compromise between perceptual impact and robustness is very hard to accomplish in this scenario.

Several speech watermarking systems that use one or a combination of the above mentioned DWM methods have been proposed in the literature, e.g., [Cheng & Sorensen 2001, Chen & Leung 2007, Celik et al. 2005, Geiser et al. 2005, Hofbauer et al. 2009, Sagi & Malah 2007, Gurijala & Deller 2007, Gurijala 2007]. Some of the more elaborate proposals, such as [Hofbauer et al. 2009] and [Sagi & Malah 2007], also account for aspects of signal equalization, synchronization as well as noise issues. The method of [Hofbauer et al. 2009], intended for analog aeronautical voice radio, embeds data at a bit rate of up to 450 bit/s by modifying the *phase* of narrowband speech signals. The proposal of [Sagi & Malah 2007], aiming at speech bandwidth extension, applies quantization based watermarking techniques in the Discrete Hartley Transform (DHT) domain. Reliable transmission of 600 bit/s is achieved over several typical telephony channels. However, the impact of strong speech compression, e.g., CELP coding, has not been evaluated.

The system of [Geiser et al. 2005] was designed for *digital* speech transmission systems and has also been tested with CELP speech coding. Here, the watermark message is embedded into a subspace of the linear prediction residual by using lattice-based quantization index modulation. It could be shown that hidden data with a rate of at least 300 bit/s can be reliably embedded in narrowband speech signals if digital *waveform* coders such as ITU-T G.711 [ITU-T 1972] or ITU-T G.726 [ITU-T 1990] are used. Yet, parametric CELP coding, which represents the state of the art method for mobile telephony, still poses a major challenge and has a severe impact on the hidden data. Reliable transmission over CELP coders has only been achieved at relatively low hidden data rates. An example for such a low-rate, but robust, speech watermarking method is introduced in [Celik et al. 2005]. It is proposed to modify the pitch period in voiced speech segments. This method has in fact been tested with a number of low-rate CELP codecs. However, the achieved (average) hidden data rate was only 3 bit/s which is not sufficient for the applications targeted in this thesis.

5.3.2 Bitstream Data Hiding (BSDH)

Algorithms for bitstream data hiding (BSDH) in speech and audio communication systems operate on the encoded bitstream. BSDH schemes have already been realized for the full spectrum of multimedia source codecs such as JPEG image coding, H.264 video coding, or MPEG-2 Advanced Audio Coding. However, there are also a number of proposals for BSDH in speech coding as summarized below. In general, the respective embedding methods are specific to the codec for which they have been designed. Typical techniques range from simple LSB substitution over requantization techniques to so called “reversible” methods. The latter exploit

residual redundancies in the source bitstream by applying entropy coding to less significant parts thereof. The freed bits can then be used to inject a hidden message while the original (coded) host signal can be fully recovered if the decoder is aware of the data hiding. BSDH also offers the possibility to embed one hidden bit in a *group* of source bits (e.g., several LSBs) by enforcing a parity constraint for the whole bit group. This way, still, only a single bit needs to be modified, but the concrete embedding position can be kept variable and the resulting average embedding distortion is reduced. Suitable parity constraints are obtained from so called *covering codes* [Cohen et al. 1997, Galand & Kabatiansky 2003]. Since, with BSDH, coding noise has no impact on the decoding of the hidden data, higher hidden data rates can be achieved than with the DWM approach. Even data hiding for low bit rate speech codecs becomes feasible to a certain extent.

The methods proposed in [Chen & Liu 2007], [Tian et al. 2009], and [Aoki 2009] use relatively straight-forward LSB substitution for the ITU-T G.723.1 codec at 6.3 kbit/s [ITU-T 1996*a*], for the ITU-T G.729 Annex A CS-ACELP speech codec at 8 kbit/s [ITU-T 1996*b*], and for the ITU-T G.711 compander with 64 kbit/s [ITU-T 1972], respectively. The latter contribution aims at bandwidth extension of the G.711 signal towards wideband speech by using hidden side information. Hidden bit rates of several 100 bit/s are reported in these proposals. Also [Shahbazi et al. 2010*a*] and [Shahbazi et al. 2010*b*] employ LSB substitution. Here, the GSM Full-Rate (FR) [ETSI 1990] and GSM Enhanced Full-Rate[2] (EFR) [ETSI 1998, Järvinen et al. 1997] codecs are used to hide data at various bit rates up to a few kbit/s. However, the speech quality is reduced, in particular for higher hidden bit rates. As an additional measure in these proposals, the actual choice of the LSB embedding positions depends on the characteristics of the current speech frame (e.g., energy or voicing). The target application is *covert voice communication* where the bitstream of the low rate MELP speech vocoder [Wang et al. 2002] is hidden within the bitstream of the EFR codec.

As another example, even the ETSI standard for tandem-free operation (TFO) [ETSI 1999] to transport, e.g., AMR or AMR-WB coded speech over the telephone core network can actually be interpreted as a simple BSDH method. In a TFO connection, the *two* LSBs of the speech samples which have been previously quantized according to ITU-T G.711 are overwritten with the respective codec bitstream. The quality impact of this operation on the G.711 coded speech is significant, but a decoding of the modified G.711 stream is normally not intended.

A BSDH method with *requantization* instead of LSB substitution has been proposed by [Licai & Shuozhong 2006] for the GSM Full-Rate codec. The idea of embedding of the hidden data in *bit groups* by applying the concept of covering codes (e.g., parity constraints, cf. [Galand & Kabatiansky 2003]) is pursued in [Ding 2004], [Liu et al. 2008] and [Xu & Yang 2009] for the ITU-T G.711 codec at

[2]The 3GPP EFR coder is virtually identical to the 3GPP AMR codec at a bit rate of 12.2 kbit/s [ETSI 2000, Ekudden et al. 1999].

64 kbit/s [ITU-T 1972], for the ITU-T G.729 codec at 8 kbit/s [ITU-T 1996*b*], and for the ITU-T G.723.1 codec at 5.3 kbit/s [ITU-T 1996*a*], respectively. Relatively high hidden rates can be achieved while maintaining a good speech quality, though at the expense of added delay in some proposals. A *reversible* data hiding method for the G.729 codec is proposed in [Li et al. 2008]. A decoder that is aware of the data hiding, can recover the *original* speech signal, i.e., there is no quality loss. The reported average hidden data rate is ca. 59 bit/s.

5.3.3 Joint Source Coding and Data Hiding (JSCDH)

There are several proposals for JSCDH in speech and audio communication systems that have appeared in the literature. In general, these methods are very specific to the concrete codec to which they are applied because the coding or quantization routines need to be modified directly. As for BSDH, coding noise does not have any impact on the decoding of the hidden data but a better tradeoff between embedding distortion and hidden bit rate is expected.

For example, the proposal of [Nishimura 2009] modifies the *pitch lag* of the 3GPP AMR codec [ETSI 2000, Ekudden et al. 1999] at bit rates of 12.2 kbit/s, 7.4 kbit/s, and 4.75 kbit/s by using a hybrid BSDH and JSCDH approach. Hidden data rates of a few 100 bit/s are shown to be achievable with a tolerable quality loss. [Xiao et al. 2008] modifies the ITU-T G.723.1 [ITU-T 1996*a*] and the iLBC [Andersen et al. 2004] coders. The data hiding is performed within the *spectral envelope parameters* of the codecs. Using a graph-based representation of the respective quantization codebooks, an optimal partitioning into two subsets is achieved, leading to a hidden data rate of 100 bit/s. Although such pseudo-random partitioning entails considerable memory-overhead, the obtained performance can be considered as an upper bound for data hiding in the spectral envelope parameters. Moreover, the performance advantage of JSCDH over BSDH could be confirmed experimentally in this study.

Another interesting JSCDH proposal has been made in [Chétry & Davies 2006]. Here, the quantizer for the *prediction residual* of the GSM Full-Rate codec at 13 kbit/s [ETSI 1990] has been modified for information embedding according to the JSCDH principle. The actual embedding has been carried out based on the principle of covering codes, in this case a convolutional code. In effect, again, a parity condition is enforced on the bitstream. Hence, the hidden data can be recovered by recomputing this parity equation in the decoder. More than 1 kbit/s of hidden data could be embedded in the bitstream of the GSM Full-Rate coder while the speech quality was affected only moderately.

A broader variety of narrowband speech codecs has been considered and compared in [Vary & Geiser 2007]. In particular, the GSM Full-Rate codec [ETSI 1990], the ITU-T G.711 compander [ITU-T 1972], the ITU-T G.726 ADPCM codec [ITU-T 1990], the ITU-T G.729 CS-ACELP codec [ITU-T 1996*b*], as well as the GSM Enhanced Full-Rate codec [ETSI 1998, Järvinen et al. 1997] have been stud-

ied. JSCDH schemes with a hidden bit rate of 600 bit/s were realized for each codec. The resulting loss in narrowband speech quality was negligible to moderate (depending on the specific codec). However, large quality *gains* could be obtained when bandwidth extension was carried out based on the hidden transmission of suitable parameters.

An application of JSCDH techniques to the *fixed codebook* of a CELP speech codec has been published by [Lu et al. 2005]. However, a basic CELP model with a Gaussian excitation codebook was assumed and the codebook partitioning method allowed overlapping partitions. Therefore, the error probability was not zero and the resulting bit rate was only 37 bit/s. The method of [Iwakiri & Matsui 1999] addresses the ITU-T G.729 CS-ACELP codec [ITU-T 1996*b*] and achieves a maximum hidden bit rate of 200 bit/s. An improved JSCDH method for state-of-the-art ACELP codecs has been proposed in [Geiser & Vary 2007*a*] and further studied in [Geiser & Vary 2008*b*] and [Vary & Geiser 2007]. Hidden bit rates up to 2 kbit/s could be achieved with negligible quality loss. These techniques are detailed and discussed in a more general scope in Sections 5.4 and 5.5.

5.3.4 Discussion

This thesis aims at a robust hidden data transmission, in particular over communication channels that involve low-rate speech coding. The three candidate technologies have particular advantages (+) and disadvantages (–) which are summarized here.

Digital Watermarking (DWM)

+ DWM as such is independent from the particular transmission system.

+ A *dedicated* auditory model or a corresponding transform can be used so that the introduced noise is kept under the perceptual threshold.

– The DWM signal $\tilde{\mathbf{x}}_{\mathrm{DWM}}$ is strongly susceptible to coding noise. An adequate tradeoff between hidden bit rate, embedding distortion, and robustness is therefore difficult to achieve. In fact, only proposals for very limited bit rates have been made in the literature that are claimed to be robust to CELP transcoding.

– As a direct consequence of the coding noise sensitivity, decoding errors may be unavoidable within the recovered hidden data.

– DWM entails both a high computational and a high design complexity. The design complexity is particularly increased because of the required detection mechanisms for the watermark signal which include frame (and possibly sample) synchronization, equalization, as well as error detection (or correction).

Bitstream Data Hiding (BSDH)

+ BSDH is immune to coding noise and the hidden bits can be perfectly recovered if there is no additional channel noise.

+ The computational complexity of BSDH schemes is, due to their conceptual simplicity, often rather low. In some cases, however, where complex covering codes and requantization techniques are applied, this might not be entirely true. However, the *decoder* side complexity is low in any case which is in contrast to DWM.

+ With BSDH, any additional synchronization and equalization mechanisms become dispensable since the respective transmission system already provides digital bits based on an equalized and synchronized input.

± For the simple (exemplary) setup of Figure 5.5, BSDH has a higher embedding distortion than DWM. However, as revealed by the literature survey in Section 5.3, with BSDH, it is much easier to achieve a good tradeoff between hidden data rate and embedding distortion for data hiding in practical speech codecs.

– With BSDH, it is not possible to exploit auditory masking effects.

– BSDH is specific to the source coding scheme which it has been designed for. Therefore, any standard transcoding unit within the transmission chain will destroy (or at least severely disturb) the hidden information. Hence, to reliably preserve the hidden bits, modified transcoding units might be required that decode and re-embed the steganographic message into the new data format.

Joint Source Coding and Data Hiding (JSCDH)

+ For JSCDH, in principle, all arguments speaking in favor of BSDH are valid as well, in particular the immunity against coding noise, the low complexity and the access to equalized and synchronized bits.

+ JSCDH provides even lower embedding distortion than BSDH since it has access to the original host signal.

+ Because of its tight interaction with the codec, auditory masking effects can be exploited by JSCDH.

+ The complexity of JSCDH is even lower than for BSDH since no dedicated embedding unit is required at the encoder side. Instead, this functionality is integrated with the encoder and only a fraction $|\mathcal{C}_m|$ of the original codebook size $|\mathcal{C}|$ has to be considered by the JSCDH routine. In fact, complexity is even reduced compared to mere source coding based on the codebook $\mathcal{C}$.

+ Because JSCDH methods are an integral part of the source encoder, there is the possibility to amend the respective codec standard with such functionality, therefore allowing a scheduled upgrade of existing communication systems on a large scale.

– As with BSDH, transcoding is not possible in JSCDH based systems without destroying or at least severely disturbing the hidden data. However, any transcoding solution for BSDH that preserves the hidden data will also be applicable to the corresponding JSCDH scheme.

For the intended application of bandwidth extension with hidden side information in low-rate CELP speech coding systems, JSCDH appears to be the most promising option. JSCDH circumvents most of the disadvantages of time or transform domain DWM and brings additional benefits over BSDH schemes by integrating the information embedding with the speech encoder. For communication systems that use waveform coding such as ITU-T G.711 [ITU-T 1972] in the classical telephone network, also BSDH or even DWM approaches can deliver satisfactory performance. For VoIP or mobile telephony systems that typically employ strong compression according to the ACELP principle, a suitable JSCDH method will be devised in the following.

5.4 JSCDH for ACELP Speech Codecs

Standardized codecs based on the ACELP (Algebraic CELP) principle [Laflamme et al. 1990, Adoul & Laflamme 1997*b*] have been very successful and are widely deployed. Perhaps the most prominent examples are the 3GPP AMR codec [ETSI 2000, Ekudden et al. 1999] which is mostly used for mobile telephony and ITU-T G.729 [ITU-T 1996*b*, Salami et al. 1998] which is a popular choice for VoIP telephony systems.

In the following, a practically viable solution for JSCDH in such ACELP speech codecs is devised. First, the principle of CELP speech coding is briefly summarized (Section 5.4.1). Then, a motivation for embedding hidden data in the indices of the ACELP codebook is given (Section 5.4.2). After a review of the respective codebook design and of the relevant coding algorithms (Section 5.4.3), a new JSCDH partitioning scheme for ACELP codebooks is described (Section 5.4.4) and the corresponding novel encoding and decoding algorithms to embed and recover the hidden data are outlined (Section 5.4.5). The practical application to standardized speech codecs is then described and evaluated in Section 5.5. A generalized scheme for JSCDH with *variable bit rate* is presented in Section 5.6.

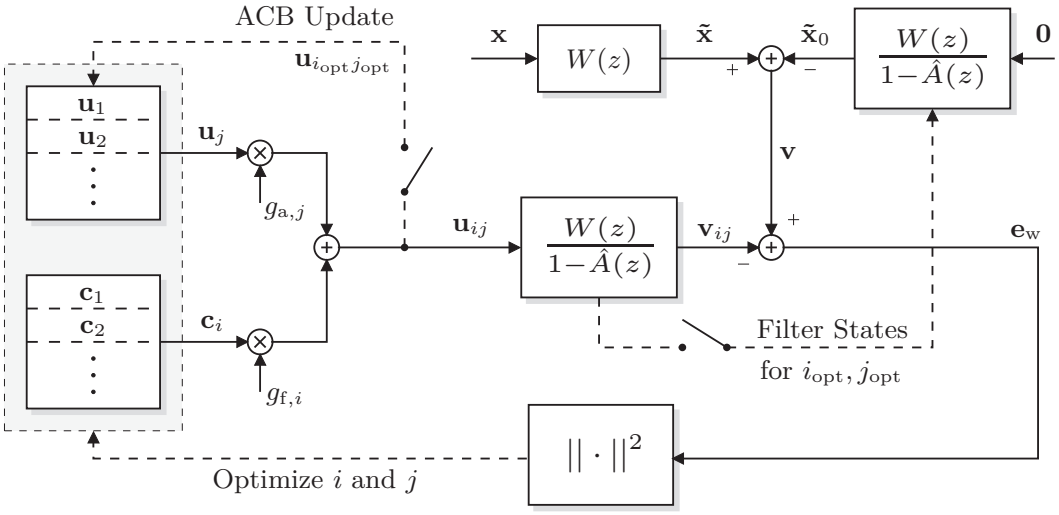


Figure 5.6: Block diagram of a CELP encoder.

5.4.1 CELP Speech Coding

The principle of *code-excited linear prediction* (CELP) speech coding has originally been proposed in [Schroeder & Atal 1985]. Comprehensive introductions are provided in numerous textbooks, e.g., [Kondoz 2004] and [Vary & Martin 2006]. Here, a brief review of the CELP principle shall be provided.

Figure 5.6 shows the block diagram of a CELP encoder with its input vector (speech frame) $\mathbf{x} \in \mathbb{R}^{L_{\text{w}}}$. First, a *short-term spectral envelope* of the current signal segment is computed using *linear predictive* (LP) analysis (e.g., [Vary & Martin 2006]). The system function of the N_{LP}-th order LP analysis filter is given as

$$1 - A(z) = 1 - \sum_{i=0}^{N_{\text{LP}}-1} a_i \, z^{-i} \tag{5.7}$$

with its coefficients a_i. The quantized version of $A(z)$ is denoted by $\hat{A}(z)$ with coefficients $\hat{a}_i$. A so called *perceptual weighting filter* for spectral shaping of the quantization noise is then derived from $A(z)$. A common choice is

$$W(z) = \frac{1 - A\left(\frac{z}{\gamma_1}\right)}{1 - A\left(\frac{z}{\gamma_2}\right)} \tag{5.8}$$

with suitable (manually tuned) constants γ_1 and γ_2. In addition, the *weighted LPC synthesis filter* is determined as

$$H(z) = \frac{W(z)}{1 - \hat{A}(z)}. \tag{5.9}$$

In the core quantization loop of the CELP encoder, *two* codebooks are used, namely the *adaptive codebook* (ACB) with its vectors $\mathbf{u}_j$ and the *fixed codebook* (FCB) with its vectors $\mathbf{c}_i$. The ACB is intended to model temporal periodicity, e.g.,

in voiced speech segments. Therefore, it is constantly updated with the previously encoded vectors $\mathbf{u}_{i_{\text{opt}}j_{\text{opt}}}$. Instead, the FCB is intended to model *stochastic* signal portions and thus contributes *innovation components* to the final candidate vector $\mathbf{u}_{ij}$. The FCB is therefore also called *stochastic* or *innovative* codebook.

The actual encoding procedure starts by weighting the input vector $\mathbf{x}$ with the filter $W(z)$ and by subtracting the zero impulse response $\tilde{\mathbf{x}}_0$ of the weighted synthesis filter $H(z)$ whereby the filter *states* are synchronized with the states of *corresponding* filter *in* the core loop that have been obtained for the *best* vector $\mathbf{u}_{ij}$ of the *previous* frame. The "filter ringing" vector $\tilde{\mathbf{x}}_0$ is subtracted from $\tilde{\mathbf{x}}$ to facilitate a filtering without memory, see, e.g., [Kondoz 2004]. The resulting vector $\mathbf{v}$ is usually called the *CELP target signal*. In the core quantization loop of the encoder, this target signal is approximated by filtered contributions that are taken from the adaptive and fixed codebooks.

The ACB and FCB indices (i and j) are optimized in a *closed quantization loop* that is based on the *analysis-by-synthesis* principle.[3] For each step of the analysis-by-synthesis procedure, one entry from the ACB $\mathbf{u}_j$ and one entry from the FCB $\mathbf{c}_i$ are individually weighted by (optimally determined) gain factors $g_{\text{a},j}$ and $g_{\text{f},i}$, cf. [Paulus 1997]. The sum of both contributions forms a candidate vector

$$\mathbf{u}_{ij} = g_{\text{a},j}\,\mathbf{u}_j + g_{\text{f},i}\,\mathbf{c}_i \tag{5.10}$$

which is then filtered through the weighted synthesis filter $H(z)$ from (5.9). Since the "filter ringing" contribution $\tilde{\mathbf{x}}_0$ has been removed from the target signal, this filtering operation can be replaced by a simple multiplication of $\mathbf{u}_{ij}$ with the lower triangular Toeplitz convolution matrix

$$\mathbf{H} = \begin{pmatrix} h(0) & 0 & \cdots & 0 \\ h(1) & h(0) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ h(L_{\text{SF}}-1) & h(L_{\text{SF}}-2) & \cdots & h(0) \end{pmatrix}, \tag{5.11}$$

where $h(k)$ is the (truncated) impulse response that corresponds to $H(z)$. The resulting candidate vector $\mathbf{v}_{ij} = \mathbf{H}\,\mathbf{u}_{ij}$ is subtracted from the target $\mathbf{v}$ and the energy of the so called *weighted error* signal $\mathbf{e}_{\text{w}} = \mathbf{v} - \mathbf{v}_{ij}$ is computed. The optimum ACB and FCB indices are then determined by minimizing the energy of the error $\mathbf{e}_{\text{w}}$:

$$(i_{\text{opt}}, j_{\text{opt}}) = \arg\min_{i,j} ||\mathbf{e}_{\text{w}}||^2. \tag{5.12}$$

As a joint optimization over all entries of the ACB and FCB would be too complex, usually a consecutive optimization of j and i (as well as $g_{\text{a},j}$ and $g_{\text{f},i}$) is performed. Here, it is assumed that, in a first step, the optimum ACB vector $\mathbf{u}_{j_{\text{opt}}}$ and the

[3] The analysis-by-synthesis procedure is usually carried out on smaller *subframes* than the LP analysis, i.e., the codevector dimension is equal to the subframe length L_{SF}.

associated gain $g_{\mathrm{a},j_{\mathrm{opt}}}$ have already been determined. Then, the target signal $\mathbf{v}$ can be updated as $\mathbf{v}_1 = \mathbf{v} - g_{\mathrm{a},j}\,\mathbf{H}\,\mathbf{u}_{j_{\mathrm{opt}}}$ and the criterion for finding the optimum FCB index i_{opt} becomes

$$i_{\mathrm{opt}} = \arg\min_i ||\mathbf{v}_1 - g_{\mathrm{f},i}\,\mathbf{H}\,\mathbf{c}_i||^2. \tag{5.13}$$

Partial differentiation of the error energy w.r.t. $g_{\mathrm{f},i}$ and equating to zero yields the optimum gain factor for each i:

$$g_{\mathrm{f},i} = \frac{\mathbf{v}_1^{\mathrm{T}}\,\mathbf{H}\,\mathbf{c}_i}{||\mathbf{H}\,\mathbf{c}_i||^2}. \tag{5.14}$$

Inserting this result, the optimization criterion (5.13) becomes

$$i_{\mathrm{opt}} = \arg\min_i ||\mathbf{v}_1||^2 - \frac{(\mathbf{v}_1^{\mathrm{T}}\,\mathbf{H}\,\mathbf{c}_i)^2}{||\mathbf{H}\,\mathbf{c}_i||^2} = \arg\max_i \frac{(\mathbf{v}_1^{\mathrm{T}}\,\mathbf{H}\,\mathbf{c}_i)^2}{||\mathbf{H}\,\mathbf{c}_i||^2} \tag{5.15}$$

since $(\mathbf{v}_1^{\mathrm{T}}\,\mathbf{H}\,\mathbf{c}_i)^2/||\mathbf{H}\,\mathbf{c}_i||^2 \geq 0$ and $\mathbf{v}_1 \equiv$ const. for the current frame. Introducing the "backward filtered target vector" $\mathbf{d}^{\mathrm{T}} = \mathbf{v}_1^{\mathrm{T}}\,\mathbf{H}$ and the "impulse response correlation matrix" $\mathbf{\Phi} = \mathbf{H}^{\mathrm{T}}\,\mathbf{H}$, (5.15) further simplifies to the final expression:

$$i_{\mathrm{opt}} = \arg\max_i \frac{(\mathbf{d}^{\mathrm{T}}\,\mathbf{c}_i)^2}{\mathbf{c}_i^{\mathrm{T}}\,\mathbf{\Phi}\,\mathbf{c}_i}. \tag{5.16}$$

The optimum gain factor $g_{\mathrm{f},i_{\mathrm{opt}}}$ is finally obtained with (5.14).

Summarizing, the following parameters are transmitted by a typical CELP encoder:

- The spectral envelope in the form of the filter coefficients $\hat{a}_i$. The quantization is often performed in another domain, e.g., using *line spectrum frequencies* (LSFs) [Itakura 1975].
- The quantized gain factors $\hat{g}_{\mathrm{f},i_{\mathrm{opt}}}$ and $\hat{g}_{\mathrm{a},j_{\mathrm{opt}}}$.
- The ACB index j_{opt}. Usually, this is represented in the form of a *pitch lag* that matches the periodicity of voiced speech. Thereby, a fractional resolution (smaller than one sample) is beneficial.
- The FCB index i_{opt}.

The spectral envelope (LSF) parameters are usually transmitted once or twice per frame, whereas the codebook indices and the gains are determined and transmitted on a subframe basis.

The CELP *decoder* is relatively simple. The optimum excitation sequence $\mathbf{u}_{i_{\mathrm{opt}}j_{\mathrm{opt}}}$ is reconstructed for each frame and the LPC synthesis filter with its system function $1/(1-\hat{A}(z))$ is applied. The additional implementation of *perceptual postfilters*, e.g., [Chen & Gersho 1995], is common.

5.4.2 Eligibility of CELP Parameters for Data Hiding

It is rather obvious that the application of JSCDH to the different CELP parameters has different, characteristic, effects on the speech quality. Also, the robustness of the hidden data may vary depending on the particular parameter type and the respective quantization mechanism. The advantages and disadvantages for data hiding in the different parameter types are discussed in the following.

Spectral Envelope (LSF Parameters)

The spectral envelope is essential in reproducing the characteristics of speech sounds. Moreover, the performance of the entire encoder depends on the accurate representation and quantization of the respective parameters because the achievable prediction gain is to a large extent determined by the spectral envelope. Usually, quantizers for, e.g., LSF parameters are strongly tuned to the source characteristics in order to achieve sufficiently low bit rates. Consequently, only comparatively low *hidden* bit rates can be expected from the application of JSCDH to the spectral envelope parameters.

However, as a potential benefit, the spectral envelope can be easily recomputed from a decoded PCM signal with sufficient accuracy and the hidden data remains recoverable under certain conditions, e.g., [Gurijala & Deller 2007, Gurijala 2007].

Codebook Gains

The gains for the adaptive and fixed codebooks can be considered inappropriate for data hiding in most cases since the respective quantizer codebooks are usually heavily optimized and the quality sensitivity is relatively high.

Pitch Lag

Like the spectral envelope parameters, also the pitch lag parameter is, on the one hand, perceptually extremely important for voiced speech segments and, on the other hand, responsible for achieving high prediction gains. Also here, it can be reasoned that only comparatively low hidden bit rates can be expected with JSCDH, at least in sensitive speech segments.

Low bit rate data hiding in the pitch lag parameter of CELP codecs has nevertheless been investigated and proposed in the literature, e.g., [Nishimura 2009], see Section 5.3.3.

Innovative Codebook Contribution

The innovative codebook contribution from the fixed codebook (FCB), here in particular the so called ACELP codebook, is much more suitable for JSCDH than the previous CELP parameters. First of all, for almost all relevant coders, the related bits form between 40% and 60% of the entire bitstream, i.e., there is, in

principle, a good potential for data hiding. Furthermore, in bit sensitivity studies that assess the quality impact of bit errors at different bitstream positions, FCB indices are often among the least significant bits, e.g., [Estrada et al. 1996] and [ETSI 2005, Table 6]. Another advantage of applying JSCDH on the FCB bits is that the FCB index for each speech frame is determined in the "innermost" component of the algorithm, i.e., no other parameters of the current frame (except the FCB gain) depend on the FCB index. In particular, the FCB index is not used for any further signal prediction in the current signal frame, i.e., any potential error propagation effects can be entirely avoided.

The above mentioned facts alone could already motivate to employ JSCDH based on the FCB. However, there is an additional distinctive advantage that can be attributed to the structure of the ACELP code. Usually, any codebook search for quantization is *exhaustive*, so that the best codevector can be determined subject to the given criterion. However, the ACELP codebook is extremely large because it is not stored explicitly but rather defined by an algebraic construction rule. For computational reasons, practical coders only implement quantization routines that perform a *by far* non-exhaustive codebook search. This fact can be elegantly exploited to improve the JSCDH performance as detailed in Section 5.4.4.

5.4.3 The ACELP Codebook

In the original proposal for CELP coding [Schroeder & Atal 1985], a random codebook of Gaussian vectors has been used to form the fixed codebook (FCB), but soon complexity and storage considerations led to structured gain-shape approaches.

A relatively simple, yet efficient codebook structure is used in so called ACELP (Algebraic CELP) codecs. The "ACELP codebook" [Laflamme et al. 1990, Adoul & Laflamme 1997*b*] is based on sparsely distributed pulse contributions and facilitates a particularly efficient determination of the innovation sequence, i.e., the FCB vector $\mathbf{c}_i$. For the ACELP codebook, elaborate (and usually non-exhaustive) codebook search algorithms have been developed that greatly reduce the computational effort to evaluate (5.16). Still, "near-optimal" solutions can be obtained.

The codevectors $\mathbf{c}$ of the ACELP codebook[4] are defined by a limited number N_P of unit pulse contributions $(+1/-1)$ that are distributed over a zero vector of length L_SF (the subframe length). Still, this sparse design alone does not reduce the computational load of the codebook search sufficiently. Therefore, to reduce the effort for the optimization of the pulse positions and signs as well as to reduce the number of bits to represent a pulse position, the allowable positions for each pulse are restricted to so called *tracks* which are, essentially, interleaved subgrids of the candidate vector to be formed.[5] The number of tracks in a concrete codebook design is denoted by N_T.

[4] The index i is omitted in the following for notational convenience.

[5] This approach is also called "Interleaved Single-Pulse Permutation" (ISPP).

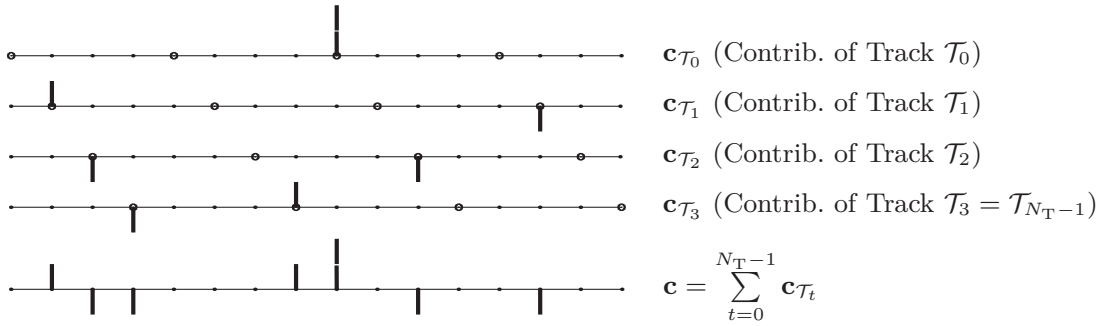


Figure 5.7: Example ACELP codevector with two pulses per track. The admissible pulse positions for each track are marked with thick black dots. ($L_\mathrm{SF} = 16$, $N_\mathrm{T} = 4$, $N_\mathrm{P} = 8$)

With the sign $s_\mathrm{p}(p_n) \in \{-1, +1\}$ of the n-th pulse and $\delta(k) = 1$ for $k = 0$ and 0 otherwise, the components of an ACELP codevector $\mathbf{c} = (c(0), c(1), \ldots, c(L_\mathrm{SF}))^\mathrm{T}$ can be written as

$$c(k) = \sum_{n=0}^{N_\mathrm{P}-1} s_\mathrm{p}(p_n) \cdot \delta(k - p_n), \quad k \in \{0, \ldots, L_\mathrm{SF} - 1\}, \tag{5.17}$$

where the N_P pulse positions p_n must be selected from a specific track, i.e., from the set $\mathcal{T}_t = \{t, N_\mathrm{T} + t, 2 \cdot N_\mathrm{T} + t, \ldots\}$ for a track with index $t \in \{0, \ldots, N_\mathrm{T} - 1\}$. In a concrete codebook design, a certain number of pulses must be allocated to each track. The construction of an example codevector with *two* pulses per track is illustrated in Figure 5.7. Also, as a practical example, the GSM EFR codec [ETSI 1998, Järvinen et al. 1997] with its subframe length of $L_\mathrm{SF} = 40$ places $N_\mathrm{P}/N_\mathrm{T} = 2$ pulses within each of its $N_\mathrm{T} = 5$ tracks. Each pulse can therefore assume one out of $L_\mathrm{SF}/N_\mathrm{T} = 8$ possible positions.

With codevectors that are composed of sparsely distributed pulses, the numerator of the CELP criterion (5.16) can be simplified as

$$C^2(\mathbf{c}) = \left(\mathbf{d}^\mathrm{T}\mathbf{c}\right)^2 = \left(\sum_{n=0}^{N_\mathrm{P}-1} s_\mathrm{p}(p_n)\, d(p_n)\right)^2. \tag{5.18}$$

where $d(p_n)$ are the respective components of the vector $\mathbf{d} = \mathbf{H}^\mathrm{T}\mathbf{v}_1$. Likewise for the denominator:

$$E(\mathbf{c}) = \mathbf{c}^\mathrm{T}\boldsymbol{\Phi}\mathbf{c} = \sum_{n=0}^{N_\mathrm{P}-1} \Phi(p_n, p_n) + 2\sum_{n=0}^{N_\mathrm{P}-2}\sum_{m=n+1}^{N_\mathrm{P}-1} s_\mathrm{p}(p_n)\, s_\mathrm{p}(p_m)\, \Phi(p_n, p_m) \tag{5.19}$$

with the elements $\Phi(p_n, p_m)$ of the matrix $\boldsymbol{\Phi} = \mathbf{H}^\mathrm{T}\mathbf{H}$. The ultimate goal is to maximize the term $C^2(\mathbf{c})/E(\mathbf{c})$ over the entire fixed codebook, i.e., for all possible combinations of pulse positions.

However, in the course of efficient ACELP codebook search algorithms, in particular when based on a depth-first search tree (see Section 5.4.5), a series of *partial* evaluations of (5.18) and (5.19) is required. In this case, only a *subset* of the N_P pulse positions is considered at a time. The successive optimization of pulse position subsets significantly reduces the number of combinatorial combinations and therefore the computational complexity of the codebook search. A "partial evaluation" of $C^2(\mathbf{c})/E(\mathbf{c})$ over the subset $\mathcal{P} \subset \{0, \ldots, N_\mathrm{P} - 1\}$ of pulses is written as $C^2_\mathcal{P}/E_\mathcal{P}$ with

$$C_\mathcal{P} = \sum_{n \in \mathcal{P}} s_\mathrm{p}(p_n)\, d(p_n) \tag{5.20}$$

and

$$E_\mathcal{P} = \sum_{n \in \mathcal{P}} \Phi(p_n, p_n) + 2 \sum_{\substack{n,m \in \mathcal{P} \\ n<m}} s_\mathrm{p}(p_n)\, s_\mathrm{p}(p_m)\, \Phi(p_n, p_m), \tag{5.21}$$

where the second summation in (5.21) is computed over all combinations of n and m with $n < m$.

Full ACELP Search

A full ACELP search can be considered impractical in almost all practically relevant situations. For instance in a codec with only $N_\mathrm{P} = 5$ pulses that may assume eight positions each, already $5 \cdot (3 + 1) = 20$ bit are needed to address the FCB vector (including the signs). This amounts to 2^{20} evaluations of $C^2(\mathbf{c})/E(\mathbf{c})$ per subframe. With a typical subframe duration of 5 ms, approx. $2.1 \cdot 10^8$ evaluations per second would be required. Therefore, it can be concluded that the complexity reduction that is achieved by the combination of the sparse codebook design with the restriction of the positions to predefined tracks is still not yet sufficient to be manageable, especially for mobile devices.

Non-Exhaustive ACELP Search Procedures

To further reduce the computational load, standardized codecs use (by far) *non-exhaustive* search strategies for the ACELP codebook. Consequently, a tradeoff between speech quality and computational complexity has to be found. Some of the more popular methods to achieve the complexity reduction are summarized in the following.

In many ACELP codecs, an initial estimate for the optimum codevector $\mathbf{c}_{i_\mathrm{opt}}$ is obtained from a so-called "pulse position likelihood vector" $\mathbf{b}$ with components $b(k)$. In the simplest case, e.g., in Annex A of ITU-T G.729 [ITU-T 1996*b*], $\mathbf{b}$ is equal to the backward filtered target vector $\mathbf{d}$, i.e., $\mathbf{b} = \mathbf{d}$. Such a heuristic target function facilitates several algorithmic simplifications. In particular, the pulse signs $s_\mathrm{p}(k)$ can be predetermined based on $\mathbf{b}$:

$$s_\mathrm{p}(k) = \mathrm{sign}\left(b(k)\right), \tag{5.22}$$

therefore decoupling the sign computation from the position optimization. The vector $\mathbf{d}$ and the matrix $\mathbf{\Phi}$ are then updated in advance to include the predetermined sign information:

$$d'(k) = s_{\mathrm{p}}(k) \cdot d(k) \tag{5.23}$$

and

$$\Phi'(k_1, k_2) = s_{\mathrm{p}}(k_1)\, s_{\mathrm{p}}(k_2) \cdot \Phi(k_1, k_2), \tag{5.24}$$

thus simplifying the computation of (5.18) and (5.19) or (5.20) and (5.21), respectively. This method is for example applied in ITU-T G.723.1 [ITU-T 1996*a*], ITU-T G.729 [ITU-T 1996*b*, Salami et al. 1998], and in the 3GPP EFR codec [ETSI 1998, Järvinen et al. 1997]. In the latter standard, moreover, the positions of two very significant pulses are predetermined by setting them to the (local) maxima of $|b(k)|$. But still, even an exhaustive search over all possible position combinations of the non-predetermined pulses is often much too complex. The EFR codec, for instance, uses 8 pulses with non-predetermined positions. Each pulse may assume 8 possible positions, i.e., $8^8 \approx 1.7 \cdot 10^7$ candidate vectors $\mathbf{c}$ would need to be evaluated in each 5 ms subframe. Consequently, at least for codecs with a higher number of pulses, a clearly non-exhaustive search procedure must be employed. This can be achieved in various ways, for example:

- A so called *focused* search approach is taken in ITU-T G.723.1 and G.729. Here, the evaluation of the fully nested optimization loops is interrupted early based on certain criteria.

- A *depth-first tree search* approach [Adoul & Laflamme 1997*a*] is pursued in the EFR codec and in Annex A of ITU-T G.729. Here, the pulses are represented as the layers of a search tree. The tree nodes represent the potential pulse positions while a path from the tree root to a tree leaf defines a complete codevector. The idea of the depth-first approach is to investigate only a few promising paths through the tree layers which are optimized locally. This is realized by successively updating the CELP criterion with contributions of small *groups of pulses* (e.g., pulse pairs), i.e., the partial summations of (5.20) and (5.21) are used. An example search tree is shown in Figure 5.8(a). This will be discussed in more detail in Section 5.4.5. The depth-first method provides a relatively high quality that is close to the optimum.[6] As possible extensions, tree pruning methods have been investigated, e.g., [Byun et al. 2002, Falahati et al. 2008].

[6] In an experiment, the standardized ITU-T G.729A codec was compared with a modified version based on a *full* ACELP search. The averaged PESQ scores [ITU-T 2001, Rix et al. 2001] for both codec versions differed by less than 0.03, indicating that they are equivalent.

- Also, several variants of ACELP codebook searches based on *pulse replacement* techniques have been proposed. Here, an initially selected codevector (possibly according to the heuristic function $\mathbf{b}$), is repeatedly updated by repositioning pulses that contribute the least to the CELP criterion, see [Park et al. 2002, Lee et al. 2003, Chen et al. 2010]. This kind of codebook search has been standardized for ITU-T G.729.1 [ITU-T 2006, Ragot et al. 2007], see [Massaloux et al. 2007]. A related method that works without any iterations has been proposed by [Lee et al. 2007]. It is particularly suited for ACELP codebooks with very few pulses.

Encoding of the ACELP Pulse Signs

If a codec places multiple pulses in one of its tracks, i.e., $N_\mathrm{P}/N_\mathrm{T} > 1$, efficient encoding schemes can be used for the pulse signs $s_\mathrm{p}(p_n)$. For example in the simplest case of two pulses per track, only a single bit needs to be spent for the two signs because the pulse position *order* in the bitstream can be used to deliver additional information. For instance if the first pulse discovered in the bitstream is placed at a position with lower index than the second pulse in the bitstream, the pulses have equal signs and unequal signs otherwise. Therefore, only a single pulse sign needs to be transmitted. Related schemes for more than two pulses per track are applied in the AMR-WB codec [ETSI 2001*b*, Bessette et al. 2002]. Naturally, the encoder has to rearrange the bits related to the pulse positions so that the desired sign information can be inferred from the bitstream order.

5.4.4 Novel ACELP Codebook Partitioning: "Algebraic Binning"

To apply the JSCDH principle to the ACELP codebook, a partitioning method must be devised based on the pulse signs and/or positions. As discussed above, signs are predetermined before the actual codebook search in most codecs. Therefore, a novel *position*-based codebook partitioning is proposed here.

The basic idea is very simple. The number of admissible positions for certain pulses is restricted to one half, one quarter, or less of the original set of pulse positions. This restriction implicitly defines the M sub-codebooks for data hiding. In the simplest case, for instance, to hide a one-bit message m in the pulse position index of a pulse that is located in the track with index t, the set of admissible positions $\mathcal{T}_t = \{t, N_\mathrm{T}+t, 2\,N_\mathrm{T}+t, \ldots\}$, is partitioned into a 0-set (for $m = 0$) such as $\mathcal{T}_t^0 = \{t, 2\,N_\mathrm{T}+t, 4\,N_\mathrm{T}+t, \ldots\}$ and a 1-set (for $m = 1$), e.g., $\mathcal{T}_t^1 = \{N_\mathrm{T}+t, 3\,N_\mathrm{T}+t, 5\,N_\mathrm{T}+t, \ldots\}$. Then, only the position set with label m must be considered during codebook search so that the hidden message may be recovered by identifying the respective set at the decoder.

However, this simple approach is not feasible in many practical codecs, in particular if multiple pulses are placed in one track. The efficient pulse sign encoding scheme (see previous section) has an undesired side effect. Since the bitstream positions of multiple pulse indices in one track may vary depending on the signs, the

decoder cannot distinguish these pulses anymore. Concretely, it becomes impossible for the decoder to identify the particular pulse position index in the bitstream that actually contains the hidden data. To resolve this ambiguity, an algebraic coupling of pulse position indices with the bits of the message m can be established. Therefore, it is assumed (for the time being), that the possible pulse positions of each track are enumerated by the natural numbers, i.e., the (binary) index i_n of a pulse at position p_n in track $\mathcal{T}_t$ is given by

$$i_n = \frac{p_n - t}{N_\mathrm{T}}. \tag{5.25}$$

The set of admissible pulse positions $\mathcal{T}_t^m$ for each message m can then be defined by an *index constraint* equation that is *symmetric* w.r.t. the ambiguous bits.

As an example, two pulses p_{n_1} and p_{n_2} in track $\mathcal{T}_t$ are considered, where data hiding shall focus on the *second* pulse p_{n_2} only, i.e., the set of admissible positions for p_{n_1} is $\mathcal{T}_t$ (unchanged) while, for p_{n_2}, the limited sets $\mathcal{T}_t^m$ are used. A possible index constraint is then given by

$$(i_{n_1} \oplus i_{n_2}) \mod M \stackrel{!}{=} m \tag{5.26}$$

where $\oplus$ is the exclusive bitwise disjunction (XOR) operator. Equation (5.26) is obviously symmetric w.r.t. the ambiguous position indices i_{n_1} and i_{n_2} (the variables are interchangeable). Solving for i_{n_2} with a given $i_{n_1} \in \mathcal{T}_t$ as well as a given $m \in \mathbb{M}$ (whereby $m \bmod M = m$) yields the admissible pulse position indices i_{n_2}. With a given subframe length L_SF, the constraint equation (5.26) has $L_\mathrm{SF}/(N_\mathrm{T} \cdot M)$ unique solutions for each message m:

$$i_{n_2} = i_{n_1} \oplus (m + \nu \cdot M) \quad \text{with} \quad \nu \in \left\{0, 1, \ldots, \frac{L_\mathrm{SF}}{N_\mathrm{T} \cdot M} - 1\right\}. \tag{5.27}$$

The number of solutions is consistent with the number of MSBs of i_{n_2} that can be interpreted as "don't care" bits w.r.t. (5.26), i.e., $\mathrm{ld}\,(L_\mathrm{SF}/N_\mathrm{T}) - \mathrm{ld}\,M$. Conversely, $\mathrm{ld}\,M$ LSBs of i_{n_2} are used to actually hide the message m. Finally, with (5.25), the admissible pulse positions p_{n_2} and hence the desired sets $\mathcal{T}_t^m$ can be obtained:

$$\mathcal{T}_t^m = \left\{ N_\mathrm{T} \cdot (i_{n_1} \oplus (m + \nu \cdot M)) + t \;\middle|\; \nu = 0, 1, \ldots, \frac{L_\mathrm{SF}}{N_\mathrm{T} \cdot M} - 1 \right\}. \tag{5.28}$$

From (5.28), it becomes obvious that, to establish the sets $\mathcal{T}_t^m$, the (candidate) index i_{n_1} must be known beforehand (i.e., $\mathcal{T}_t^m = \mathcal{T}_t^m(i_{n_1})$). A suitable organization of the codebook search can ensure this property as will become clear in the following section.

The *decoding* of the hidden message m is achieved by simply recomputing the constraint equation (5.26) based on the *received* pulse position indices. Thereby, the symmetry property of (5.26) elegantly resolves the ambiguity of the received indices i_{n_1} and i_{n_2} w.r.t. their order in the bitstream.

Numerical Example for Algebraic Binning

The proposed codebook partitioning method shall now be illustrated with a concrete example. An ACELP codebook for subframes of length $L_{\mathrm{SF}} = 12$ with $N_{\mathrm{P}} = 6$ pulses in $N_{\mathrm{T}} = 3$ tracks (two pulses per track) is considered. Each pulse may therefore assume one out of $L_{\mathrm{SF}}/N_{\mathrm{T}} = 4$ possible positions. The data hiding shall focus on the *second* pulse in track $\mathcal{T}_1 = \{1, 4, 7, 10\}$. The *first* pulse p_{n_1} in this track has already been selected, e.g., $p_{n_1} = 7$. According to (5.25), the corresponding index is $i_{n_1} = (7-1)/3 = 2$. A message $m = 1$ out of $M = 2$ possible messages shall now be transmitted. Consequently, there are $12/(3 \cdot 2) = 2$ unique solutions for (5.27):

$$i_{n_2} = \begin{cases} 2 \oplus (1 + 0 \cdot 2) \\ 2 \oplus (1 + 1 \cdot 2) \end{cases} = \begin{cases} 3 \\ 1 \end{cases}$$

which, using (5.25), translate into the admissible pulse positions:

$$p_{n_2} = \begin{cases} 3 \cdot 3 + 1 \\ 1 \cdot 3 + 1 \end{cases} = \begin{cases} 10 \\ 4 \end{cases},$$

i.e., the restricted pulse position set is $\mathcal{T}_1^1 = \{4, 10\}$. If, for example, the pulse position $p_{n_2} = 4$ is selected, the corresponding index $i_{n_2} = 1$ is transmitted and the hidden message can be decoded using (5.26):

$$\hat{m} \;=\; (2 \oplus 1) \bmod 2 \;=\; 3 \bmod 2 \;=\; 1 \;=\; m.$$

For $p_{n_2} = 10$, i.e., $i_{n_2} = 3$, the same result is obtained:

$$\hat{m} \;=\; (2 \oplus 3) \bmod 2 \;=\; 1 \bmod 2 \;=\; 1 \;=\; m.$$

An analogous example can be constructed for $m = 0$ leading to the complementary restricted pulse position set $\mathcal{T}_1^0 = \{1, 7\}$.

Computation of the Hidden Bit Rate

The described partitioning procedure can be applied for more than one pulse pair and for more than one track, facilitating higher hidden bit rates and a complete partitioning of the ACELP codebook. In total, the hidden bit rate R_{DH} depends on the number of pulses used for data hiding and on the ratio $|\mathcal{T}_{t_j}|/|\mathcal{T}_{t_j}^{m_j}|$ for each pulse. It can be computed as

$$R_{\mathrm{DH}} = \frac{f_{\mathrm{s}}}{L_{\mathrm{SF}}} \sum_{j=0}^{N_{\mathrm{P}}-1} \mathrm{ld}\, \frac{|\mathcal{T}_{t_j}|}{|\mathcal{T}_{t_j}^{m_j}|}. \tag{5.29}$$

Naturally, with the described scheme, the effective codebook size for each subframe is reduced by R_{DH} bits to $N_{\mathrm{P}} \cdot \mathrm{ld}\,(L_{\mathrm{SF}}/N_{\mathrm{T}}) - R_{\mathrm{DH}}$ bits.

The Particular Benefit of ACELP JSCDH

As mentioned, ACELP codebook search algorithms are by far non-exhaustive and the vast majority of codebook entries is not examined in standard search procedures. This fact is elegantly exploited in the present JSCDH scheme.

With the proposed codebook partitioning, usually, each of the $2^{R_{\mathrm{DH}}}$ sub-codebooks will, individually, be large enough to cover an adequate portion of the full ACELP codebook, i.e., even the *reduced* sub-codebooks comprise *more* entries than the codebook portion that is examined in the *standardized* search algorithm. Hence, with the present partitioning scheme, codebook entries can be taken into account that have been disregarded in the standard (non-steganographic) implementation. If it is possible to find M "equally good" sub-codebooks, each yielding a coding performance that is comparable to that of the original algorithm, then the data hiding procedure does not (or only insignificantly) degrade the resulting speech quality. In Section 5.5.3, it will be shown that this can actually be achieved with the devised JSCDH method, even for relatively high hidden bit rates.

5.4.5 Steganographic Codebook Search Algorithms

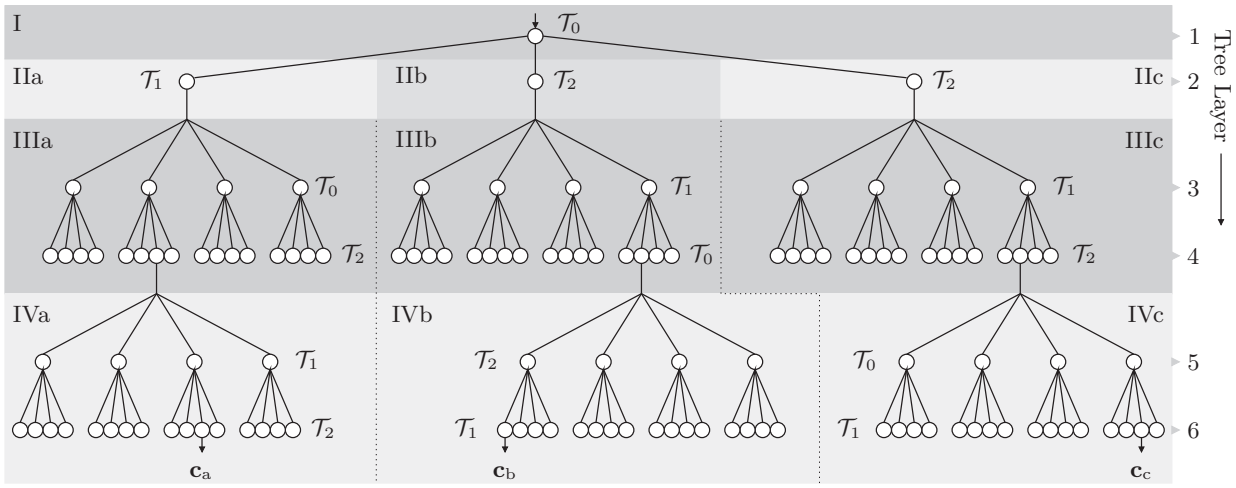
Based on a given codebook partitioning, a suitable codebook search algorithm can be developed. The proposed *steganographic* codebook search algorithms are based on the *depth-first tree search* method as briefly introduced in Section 5.4.3. The depth-first approach is known to perform close to the optimum and is widely used in standardized codecs. Naturally, also steganographic versions of other codebook search approaches (e.g., based on pulse replacement) are conceivable. Here, the underlying ideas and principles of the proposed algorithms shall be summarized. *Concrete* realizations are detailed in Section 5.5.

Exemplary Depth-First Tree Search (Standard Algorithm)

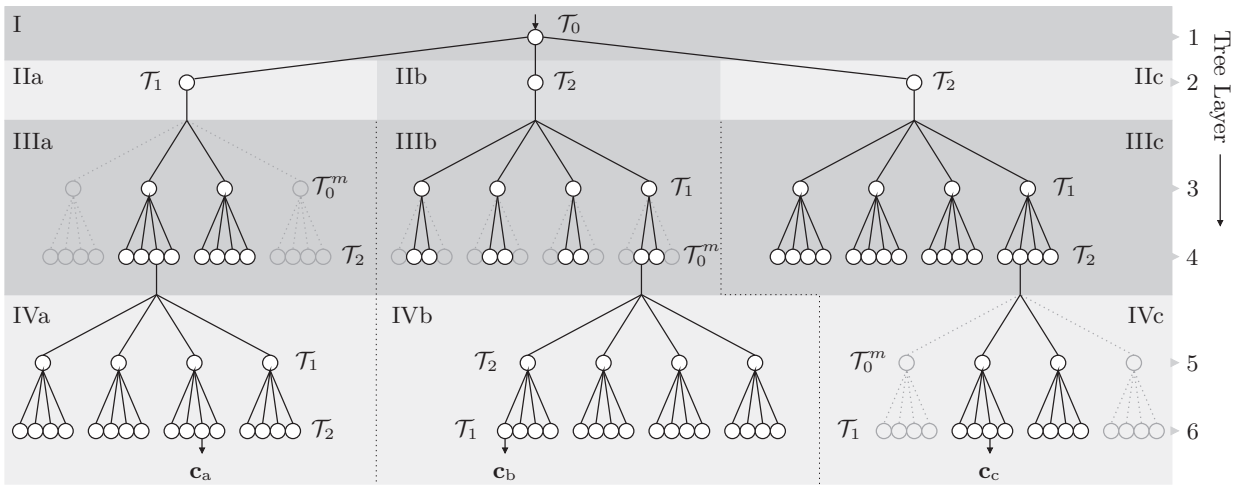
Before the actual steganographic codebook search algorithms are introduced, the principles of the *standard* depth-first tree search shall be explained in detail.

An example of a typical depth-first FCB search tree is shown in Figure 5.8(a). The (sub-)frame length is, again, $L_{\mathrm{SF}} = 12$ with $N_{\mathrm{P}} = 6$ pulses in $N_{\mathrm{T}} = 3$ tracks (two pulses per track). With each of the six tree *layers* ($N_{\mathrm{P}} = 6$), a pulse is tentatively added to the current codevector candidate. The pulses are positioned on the tracks $\mathcal{T}_0$, $\mathcal{T}_1$ or $\mathcal{T}_2$ as annotated in the figure. Each pulse may assume one out of $L_{\mathrm{SF}}/N_{\mathrm{T}} = 4$ possible positions, i.e., each tree *node* has four *child nodes*.[7] Assuming predetermined pulse signs $s_{\mathrm{p}}(p_n)$, the *full* search tree for the pulse *positions* would comprise $N_{\mathrm{L}} = 4^6 = 4096$ leafs (codevector realizations) and $N_{\mathrm{N}} = \sum_{i=1}^{6} 4^i = 5460$ nodes (intermediate processing steps). The depth-first search algorithm reduces these numbers significantly.

[7] The *root node*, as an exception, has only *three* children, corresponding to the number of algorithm *iterations* as explained below.



(a) Standard search algorithm ($R_\mathrm{DH} = 0$, $N_\mathrm{L} = 93$, $N_\mathrm{N} = 124$)



(b) Restricted search algorithm ($R_\mathrm{DH} = 1\,\mathrm{bit/vector}$, $m = 0$, $N_\mathrm{L} = 69$, $N_\mathrm{N} = 96$)

Figure 5.8: Tree representation of depth-first ACELP codebook search algorithms ($N_\mathrm{P} = 6$, $N_\mathrm{T} = 3$, track length $L_\mathrm{SF}/N_\mathrm{T} = 4$).

The exemplary depth-first FCB search proceeds in three iterations, corresponding to three "sub-trees." The outcome of each iteration is a candidate vector ($\mathbf{c}_\mathrm{a}$, $\mathbf{c}_\mathrm{b}$, and $\mathbf{c}_\mathrm{c}$) with $N_\mathrm{P} = 6$ pulses each. Ultimately, the best vector among these three candidates is selected.

In the following, it is assumed that pulse positions p_0 and p_3 are located in track $\mathcal{T}_0$. Likewise, $p_1, p_4 \in \mathcal{T}_1$ and $p_2, p_5 \in \mathcal{T}_2$. The search algorithm begins with a fixed first pulse that is positioned at the maximum magnitude of the likelihood vector $\mathbf{b} = (b(0), b(1), \ldots, b(L_\mathrm{SF} - 1))^\mathrm{T}$. The track which is associated with the maximum position $\hat{p}$ is denoted by $\mathcal{T}_{\hat{t}}$ with $\hat{t} \in \{0, 1, \ldots, N_\mathrm{T} - 1\}$. The respective pulse position $p_{\hat{t}}$ is then formally given by

$$p_{\hat{t}} = \hat{p} \qquad \text{with} \qquad \hat{p} = \arg \max_{0 \le k \le L_\mathrm{SF}-1} |b(k)| \quad \text{and} \quad \hat{t} = \hat{p} \bmod N_\mathrm{T}. \tag{5.30}$$

For the present example it is further assumed that $\hat{p}$ is located in track $\mathcal{T}_0$ (i.e., $\hat{t} = 0$) as indicated in Block I of the Figure 5.8(a). Then, three iterations are carried out, i.e., the three different sub-trees are examined:

1. $\mathcal{T}_1, (\mathcal{T}_0, \mathcal{T}_2), (\mathcal{T}_1, \mathcal{T}_2)$ — First sub-tree, Blocks IIa – IVa
2. $\mathcal{T}_2, (\mathcal{T}_1, \mathcal{T}_0), (\mathcal{T}_2, \mathcal{T}_1)$ — Second sub-tree, Blocks IIb – IVb
3. $\mathcal{T}_2, (\mathcal{T}_1, \mathcal{T}_2), (\mathcal{T}_0, \mathcal{T}_1)$ — Third sub-tree, Blocks IIc – IVc

For each iteration (sub-tree), a different assignment of the tracks to the tree layers is used which is based on a fixed permutation scheme. The braces in the track sequence denote the *full* optimization (over all possible position combinations) of a pulse group (pair). This is detailed below. Note that, in all three cases, the pulse of the second tree layer, i.e., the *first one* of the respective track sequence, must not be located in the same track as $\hat{p}$ (here $\mathcal{T}_0$).

The *first* iteration (Blocks IIa – IVa) of the algorithm shall now be described in more detail. In Block IIa, the second pulse ($p_1 \in \mathcal{T}_1$) is placed on the *local* maximum of $|b(k)|$ within track $\mathcal{T}_1$:

$$p_1 = \arg \max_{k \in \mathcal{T}_1} |b(k)|. \tag{5.31}$$

With the first two pulse positions known, Equations (5.20) and (5.21) are used to compute and store the values of $C_{\mathcal{P}_{\mathrm{IIa}}}$ and $E_{\mathcal{P}_{\mathrm{IIa}}}$ based on the subset $\mathcal{P}_{\mathrm{IIa}} = \{p_0, p_1\}$ of pulse positions.

Then, the actual depth-first search begins by optimizing pulse *pairs*. All position combinations for the two pulses $p_3 \in \mathcal{T}_0$ and $p_2 \in \mathcal{T}_2$ are examined exhaustively (Block IIIa) by maximizing the (partial) CELP criterion $C^2_{\mathcal{P}_{\mathrm{IIIa}}}/E_{\mathcal{P}_{\mathrm{IIIa}}}$ based on the pulse subset $\mathcal{P}_{\mathrm{IIIa}} = \{p_0, p_1, p'_2, p'_3\}$:

$$(p_2, p_3) = \arg \max_{\substack{p'_2 \in \mathcal{T}_2 \\ p'_3 \in \mathcal{T}_0}} \frac{C^2_{\mathcal{P}_{\mathrm{IIIa}}}}{E_{\mathcal{P}_{\mathrm{IIIa}}}}. \tag{5.32}$$

The previously stored values of $C_{\mathcal{P}_{\mathrm{IIa}}}$ and $E_{\mathcal{P}_{\mathrm{IIa}}}$ can be used for an efficient implementation of (5.32). Correspondingly, the values $C_{\mathcal{P}_{\mathrm{IIIa}}}$ and $E_{\mathcal{P}_{\mathrm{IIIa}}}$ that have been found for the best position combination (p_2, p_3) are stored for later use. In Block IVa of the figure, the final pulse pair $p_4 \in \mathcal{T}_1$ and $p_5 \in \mathcal{T}_2$ is added to the codevector by using the same procedure:

$$(p_4, p_5) = \arg \max_{\substack{p'_4 \in \mathcal{T}_1 \\ p'_5 \in \mathcal{T}_2}} \frac{C^2_{\mathcal{P}_{\mathrm{IVa}}}}{E_{\mathcal{P}_{\mathrm{IVa}}}}, \tag{5.33}$$

where $\mathcal{P}_{\mathrm{IVa}} = \{p_0, p_1, p_2, p_3, p'_4, p'_5\}$. Again, the stored values of $C_{\mathcal{P}_{\mathrm{IIIa}}}$ and $E_{\mathcal{P}_{\mathrm{IIIa}}}$ can be used for an efficient implementation of (5.33) and the values $C_{\mathcal{P}_{\mathrm{IVa}}}$ and

$E_{\mathcal{P}_{\mathrm{IVa}}}$ that have been found for the best position combination (p_4, p_5) are stored. The outcome of the first algorithm iteration is a candidate codevector $\mathbf{c}_\mathrm{a}$ with $c_\mathrm{a}(k) = \sum_{n=0}^{N_\mathrm{P}-1} s_\mathrm{p}(p_n) \cdot \delta(k - p_n)$.

The second and third iterations of the tree search algorithm work analogously: In the *second* iteration, i.e., in Block IIb of the figure, the local maximum of $\mathcal{T}_2$ is selected and assigned to p_2. In Blocks IIIb and IVb, the pulse pairs $p_1 \in \mathcal{T}_1$ and $p_3 \in \mathcal{T}_0$ as well as $p_5 \in \mathcal{T}_2$ and $p_4 \in \mathcal{T}_1$ are examined and a candidate $\mathbf{c}_\mathrm{b}$ is produced. Likewise, in the third iteration, the local maximum of $\mathcal{T}_2$ is selected and assigned to p_2. Then, the pulse pairs $p_1 \in \mathcal{T}_1$ and $p_5 \in \mathcal{T}_2$ (Block IIIc) as well as $p_3 \in \mathcal{T}_0$ and $p_4 \in \mathcal{T}_1$ (Block IVc) are examined and a candidate $\mathbf{c}_\mathrm{c}$ is produced.

Finally, the best vector among the three candidates $\mathbf{c}_\mathrm{a}, \mathbf{c}_\mathrm{b}, \mathbf{c}_\mathrm{c}$ is selected:

$$\mathbf{c}_{i_\mathrm{opt}} = \arg \max_{\mathbf{c} \in \{\mathbf{c}_\mathrm{a}, \mathbf{c}_\mathrm{b}, \mathbf{c}_\mathrm{c}\}} \frac{C^2(\mathbf{c})}{E(\mathbf{c})}. \tag{5.34}$$

Aiming at an efficient implementation, the respective values of $C(\mathbf{c})$ and $E(\mathbf{c})$ are in fact equal to the values of $C_{\mathcal{P}_{\mathrm{IV}\{\mathrm{a,b,c}\}}}$, $E_{\mathcal{P}_{\mathrm{IV}\{\mathrm{a,b,c}\}}}$ that have been previously stored in the respective iterations. In total, $N_\mathrm{L} = 93$ leafs and $N_\mathrm{N} = 124$ tree nodes have been visited to determine the final codevector.

Steganography with Restricted Depth-First Tree Search

The simplest method to realize a steganographic codebook search is to retain the standard search algorithm of the respective codec. The only required modification is to reject "illegal" codevectors in the course of the codebook search, cf. [Iwakiri & Matsui 1999]. A codevector can be considered illegal if it comprises pulse positions that are not contained in the respective sets $\mathcal{T}_t^m$ of (5.28). As a more efficient alternative, a *restricted search tree* can be established as explained in the following. However, in both cases, the covered portion of the search space is considerably reduced with each hidden bit. This, inevitably, leads to inferior speech quality.

An example of this (simple) approach is shown in Figure 5.8(b) which essentially represents the same tree as in Figure 5.8(a), but now a one-bit message $m \in \{0, 1\}$ shall be hidden in the final codevector ($R_\mathrm{DH} = 1\,\mathrm{bit/vector}$). More concretely, data hiding shall be performed on the position index p_3 which represents the *second* pulse of track $\mathcal{T}_0$. Therefore, the set $\mathcal{T}_0$ is partitioned into the subsets $\mathcal{T}_0^0$ and $\mathcal{T}_0^1$. For the present example, $m = 0$ is assumed. A complementary search tree exists for $m = 1$ which is not shown in the figure. To realize the data hiding functionality, algorithmic modifications must be applied in these specific tree layers that are associated with pulse $p_3 \in \mathcal{T}_0$. In these layers, $\mathcal{T}_0$ is simply replaced by the restricted sets $\mathcal{T}_0^m$. For the three iterations of the search algorithm, the concrete modifications can be found in Blocks IIIa (tree layer 3), IIIb (tree layer 4), and IVc (tree layer 5) where now only two instead of four possible nodes may be selected. The "illegal" tree paths, (corresponding to $\mathcal{T}_0 \setminus \mathcal{T}_0^0$) have been grayed out in the figure. If, moreover, an efficient sign encoding is used and the pulse position indices

are therefore ambiguous, it is important that the *first* pulse of this track has been positioned beforehand (in Block I) so that the set $\mathcal{T}_0^m$ can actually be computed according to (5.28).

The described steganographic version of the search tree comprises $N_\mathrm{L} = 72$ leafs and $N_\mathrm{N} = 96$ nodes, irrespective of the message m. However, as the covered portion of the search space is considerably smaller than in the non-steganographic algorithm (96 vs. 124 nodes), the speech quality is reduced (e.g., the candidate vector $\mathbf{c}_\mathrm{c}$ in Block IVc is different from the original outcome). Therefore, a better approach is to find a new tradeoff between speech quality and computational complexity for a given codebook partitioning, hence dedicated *steganographic codebook search algorithms* should be designed as shown in the following.

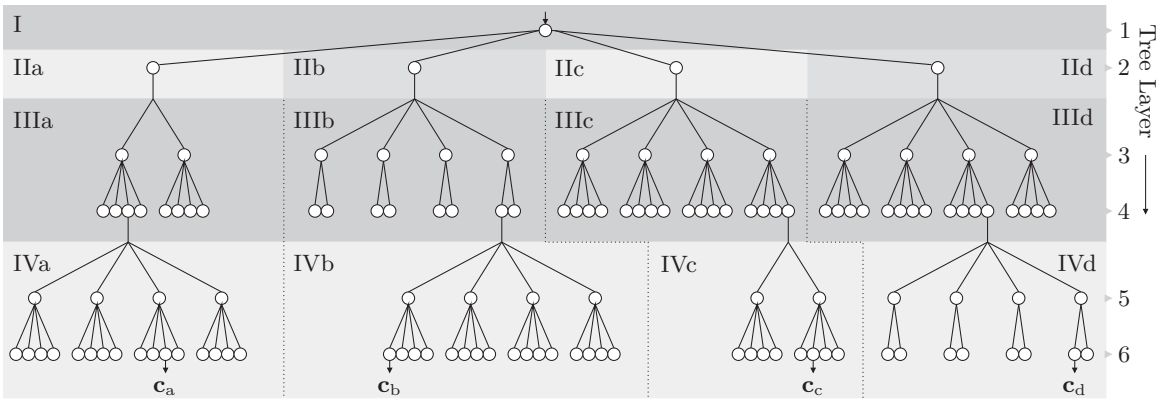
Steganographic Codebook Search with Additional Iterations

As one simple measure to re-expand the search tree, more iterations can be carried out, thus examining more sub-trees and producing *more* candidate codevectors. Hence, there is a higher probability to find a "good" codevector which yields a high speech quality. An example is depicted in Figure 5.9(a) where, compared to Figure 5.8(b), a fourth iteration has been added (IId – IVd) so that four instead of three candidate codevectors ($\mathbf{c}_\mathrm{a}$, $\mathbf{c}_\mathrm{b}$, $\mathbf{c}_\mathrm{c}$, $\mathbf{c}_\mathrm{d}$) are produced. The number of tree leafs and nodes are increased to $N_\mathrm{L} = 92$ and $N_\mathrm{N} = 129$ again, irrespective of the message m. This method is, however, limited by the number of reasonable track sequences (permutations) for the iterations (IIa – IId). Therefore, as an alternative, less pulse positions may be predetermined based on the likelihood vector $\mathbf{b}$.

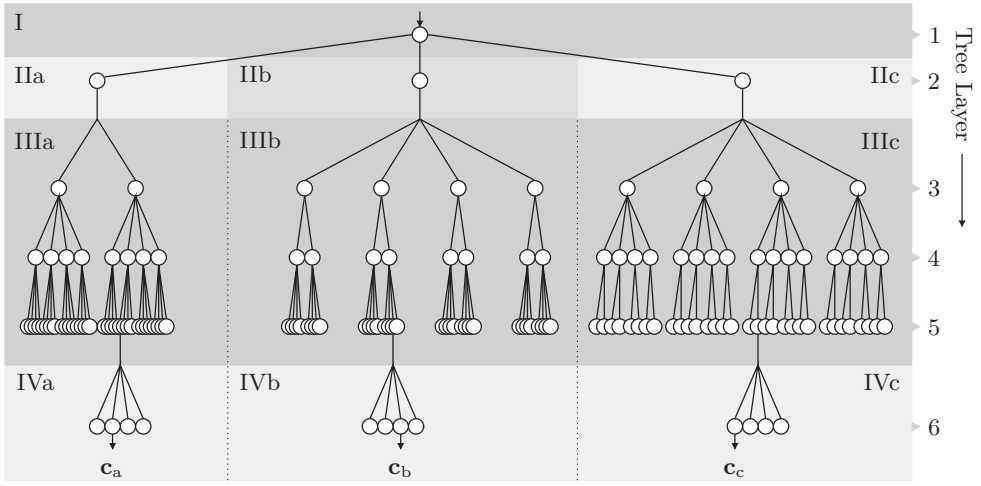
Steganographic Codebook Search with Regrouped Pulses

The depth-first search algorithms discussed so far only optimize pulse *pairs* which are successively added to the codevector. Another possibility to re-expand the search tree is to jointly process more than two pulses, i.e., pulse *triples* or *quadruples* may be considered. The concrete (local) optimization procedure as, e.g., in (5.32) has to be adapted accordingly. The result is a different algorithmic scheduling and a broader coverage of the search tree.

In the reorganized search tree of Figure 5.9(b), the joint optimization of two pulse pairs (see IIIa – IIIc and IVa – IVc of Figure 5.8(b)) has been replaced by the optimization of a pulse *triple* (IIIa – IIIc) followed by the optimization of a single pulse (IVa – IVc). The optimization procedure for the first and the second pulse (I and IIa – IIc) is unchanged from the standard algorithm and the data hiding is still conducted in tree layers 3 (Block IIIa), 4 (Block IIIb), and 5 (Block IIIc) for the three iterations. The described modifications lead to $N_\mathrm{L} = 105$ leafs and $N_\mathrm{N} = 154$ visited tree nodes which is, though a little bit more complex than the original algorithm ($N_\mathrm{N} = 124$), still in the same order of magnitude and much smaller than an exhaustive search.



(a) Steganographic codebook search with additional iteration (Blocks IId – IVd) ($R_{\mathrm{DH}} = 1\,\mathrm{bit/vector}$, $m = 0$, $N_{\mathrm{L}} = 92$, $N_{\mathrm{N}} = 129$)



(b) Steganographic codebook search with pulse regrouping: pulse triple (Blocks IIIa – IIIc) and single pulse (Blocks IVa – IVc) ($R_{\mathrm{DH}} = 1\,\mathrm{bit/vector}$, $m = 0$, $N_{\mathrm{L}} = 105$, $N_{\mathrm{N}} = 154$)

Figure 5.9: Modified depth-first ACELP codebook search algorithms.

Precedence Relations and Complexity Considerations

In any case, with the reorganization of the codebook search, also the determination of the restricted position sets $\mathcal{T}_t^m$ has to be accounted for. As the admissible positions may depend on other, previously determined pulse positions or indices, certain precedence relations have to be regarded in the algorithm design.

As an example, the restricted sets $\mathcal{T}_0^m$ shall be used to determine the pulse position $p_3 \in \mathcal{T}_0^m$. However, as in (5.28), $\mathcal{T}_0^m$ itself depends on the index i_0 of the pulse at position $p_0 \in \mathcal{T}_0$. Then, the determination of p_0 (and therefore i_0) must *always* be scheduled to a lower tree layer (i.e., closer to the tree root) than the determination of p_3 so that the value for i_0 is known before $\mathcal{T}_0^m$ must be computed.

The general idea is to optimize the unrestricted pulses (sets $\mathcal{T}_t$) *before* the restricted pulse sets $\mathcal{T}_t^m$ are computed and used. However, this should only be done if required, because an early scheduling (low tree layer) of tracks with many candidate positions considerably increases the number of tree nodes and, hence, the complexity. Moreover, a joint optimization of two or more pulses that (mutually) depend on each other should be avoided if possible because the respective sets $\mathcal{T}_t^m$ must be computed *within* a nested loop in this case which is also computationally complex.

5.5 Practical Examples

The method for JSCDH in ACELP codecs as introduced above shall now be applied to standardized and widely deployed codecs, namely the ITU-T G.729 (Section 5.5.1) and the 3GPP EFR codec which is virtually identical with the 12.2 kbit/s mode of the 3GPP AMR codec (Section 5.5.2). Codec-specific details as well as some concrete implementation aspects of the data hiding algorithms are described, cf. [Geiser & Vary 2007*a*, Vary & Geiser 2007, Geiser & Vary 2008*b*]. Then, the resulting performance is analyzed by measuring the incurred speech quality loss (Section 5.5.3). The respective quality scores are obtained with the objective PESQ evaluation tool [ITU-T 2001, Rix et al. 2001]. Moreover, objective SNR measurements as well as the results of a subjective ABX test are stated. For reference, the results of the proposed algorithms are compared with a number of other proposals from the literature (cf. Section 5.3).

5.5.1 ITU-T G.729 Annex A (CS-ACELP)

As a first (simple) example, JSCDH is applied to the fixed codebook of the ITU-T G.729 CS-ACELP codec [ITU-T 1996*b*, Salami et al. 1998]. This narrowband codec operates at a sampling frequency of $f_\mathrm{s} = 8\,\mathrm{kHz}$ and uses 5 ms subframes ($L_\mathrm{SF} = 40$) for the FCB search. In particular, Annex A of the standard (G.729A) is employed here as a reference since an efficient depth-first tree search approach is used therein to determine the positions of in total four pulses ($N_\mathrm{P} = 4$). The pulse signs $s_\mathrm{p}(p_n)$ are predetermined out of the loop as explained in Section 5.4.3. Since not more than one pulse can be placed in one track, no efficient sign encoding scheme is needed. Therefore, a pulse position index i_n is directly associated with a specific position in the bitstream and no ambiguities occur.

The FCB structure with all allowable pulse positions is tabulated in Table 5.1. A particular characteristic of the G.729 FCB is that $N_\mathrm{P} = 4$ pulses are placed in $N_\mathrm{T} = 5$ tracks. As shown in Table 5.1, this is solved by selecting p_3 either from track $\mathcal{T}_3$ *or* from track $\mathcal{T}_4$. With the given codebook structure, the full search tree to evaluate all possible position combinations has $N_\mathrm{L} = 2^{3+3+3+4} = 8192$ leafs and $N_\mathrm{N} = 8776$ nodes.

Table 5.1: FCB structure of the ITU-T G.729 codec.

pulse number (n)	valid pulse positions (p_n) and associated tracks ($\mathcal{T}_t$)
0	$0, 5, 10, 15, 20, 25, 30, 35 \quad (\mathcal{T}_0)$
1	$1, 6, 11, 16, 21, 26, 31, 36 \quad (\mathcal{T}_1)$
2	$2, 7, 12, 17, 22, 27, 32, 37 \quad (\mathcal{T}_2)$
3	$\underbrace{3, 8, 13, 18, 23, 28, 33, 38}_{\mathcal{T}_3}, \underbrace{4, 9, 14, 19, 24, 29, 34, 39}_{\mathcal{T}_4}$

Standard Codebook Search

In Annex A of G.729, the full search tree is reduced by the preselection of two promising out of eight possible positions for the first pulse and by using a depth first search approach. The preselection is, again, based on the maximum magnitudes of a pulse position likelihood vector $\mathbf{b}$ which, in this case, is simply identical with the vector $\mathbf{d}$ from the numerator of the CELP criterion (5.16). The resulting reduced tracks (containing only the two pulse positions where $|b(k)|$ is locally maximum) are denoted by $\mathcal{T}_t'$. The standard defines a codebook search with four iterations and the following *fixed* scheduling of track sequences:

1. $(\mathcal{T}_2', \mathcal{T}_3), (\mathcal{T}_0, \mathcal{T}_1)$
2. $(\mathcal{T}_2', \mathcal{T}_4), (\mathcal{T}_0, \mathcal{T}_1)$
3. $(\mathcal{T}_3', \mathcal{T}_0), (\mathcal{T}_1, \mathcal{T}_2)$
4. $(\mathcal{T}_4', \mathcal{T}_0), (\mathcal{T}_1, \mathcal{T}_2)$

The search is organized such that, in each iteration, two pulse *pairs* are successively optimized. This is indicated by the braces in the track sequences. The search tree that is spanned by this scheduling comprises $N_\mathrm{L} = 316$ leafs and $N_\mathrm{N} = 360$ nodes which poses a tremendous reduction of the computational complexity in comparison to the full search tree.

Codebook Partitioning for Steganography

As a concrete example for ACELP JSCDH in G.729A, three bits shall be hidden inside the FCB index for each subframe. This amounts to a hidden data rate of $R_\mathrm{DH} = 3\,\mathrm{bits}/5\,\mathrm{ms} = 600\,\mathrm{bit/s}$. The three message bits $m_j \in \{0, 1\}$ with $j \in \{0, 1, 2\}$ in each subframe are *individually* hidden in the three pulse position indices i_1, i_2, and i_3 that are associated with the pulse positions p_1, p_2, and p_3. As, in G.729, there is only one (or no) pulse per track, i.e., there is no ambiguity in the bitstream due to sign encoding, a relatively simple codebook partitioning scheme can be employed as shown in Table 5.2 and in Figure 5.10. The decoding of m_j is also straight forward.

Table 5.2: Restricted FCB structure for ITU-T G.729, $R_{\mathrm{DH}} = 600\,\mathrm{bit/s}$.

pulse number (n)	valid pulse positions (p_n) and associated tracks ($\mathcal{T}_t^{m_j}$)
0	$0, 5, 10, 15, 20, 25, 30, 35 \quad (\mathcal{T}_0)$
1	$1 + N_{\mathrm{T}}\, m_0,\ 11 + N_{\mathrm{T}}\, m_0,\ 21 + N_{\mathrm{T}}\, m_0,\ 31 + N_{\mathrm{T}}\, m_0 \quad (\mathcal{T}_1^{m_0})$
2	$2 + N_{\mathrm{T}}\, m_1,\ 12 + N_{\mathrm{T}}\, m_1,\ 22 + N_{\mathrm{T}}\, m_1,\ 32 + N_{\mathrm{T}}\, m_1 \quad (\mathcal{T}_2^{m_1})$
3	$3 + N_{\mathrm{T}}\, m_2,\ 13 + N_{\mathrm{T}}\, m_2,\ 23 + N_{\mathrm{T}}\, m_2,\ 33 + N_{\mathrm{T}}\, m_2, \quad (\mathcal{T}_3^{m_2})$
	$9 - N_{\mathrm{T}}\, m_2,\ 19 - N_{\mathrm{T}}\, m_2,\ 29 - N_{\mathrm{T}}\, m_2,\ 39 - N_{\mathrm{T}}\, m_2 \quad (\mathcal{T}_4^{m_2})$

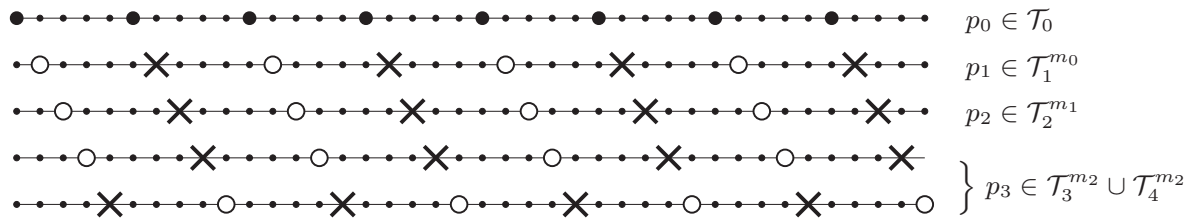


Figure 5.10: Codebook partitioning for ITU-T G.729A according to Table 5.2. The first track is unrestricted (positions marked with big black dots may be chosen). In the other tracks, positions marked with a circle indicate $m_j = 0$, positions marked with a cross indicate $m_j = 1$. ($N_{\mathrm{T}} = 5$)

Steganographic Codebook Search

Based on the chosen codebook partitioning, a novel steganographic codebook search algorithm has been devised. It must be noted that, in this case, the standardized algorithm of G.729A is not applicable because the preselection of two candidate positions according to the vector $\mathbf{b}$ may contradict with the restricted pulse positions sets in some cases, i.e., $\mathcal{T}_t' \cap \mathcal{T}_t^{m_j} = \emptyset$.

Here, based on the given hidden bit rate and codebook partitioning, the speech quality shall be *maximized*. In fact, this is easily achieved with a full tree search over the restricted codebook of Table 5.2. Using the notation from above, just a single track sequence $(\mathcal{T}_1^{m_0}, \mathcal{T}_2^{m_1}, \mathcal{T}_3^{m_2} \cup \mathcal{T}_4^{m_2}, \mathcal{T}_0)$ is evaluated. The number of visited tree leafs and nodes in this case is $N_{\mathrm{L}} = 1024$ and $N_{\mathrm{N}} = 1172$, respectively. The complexity is clearly higher than with the standard codebook search algorithm but, nevertheless, still considerably lower than a full codebook search based on the unrestricted codebook of Table 5.1 (factor of $2^3 = 8$ in terms of tree leafs).

Table 5.3: FCB structure of the 3GPP EFR codec.

pulse number (n)	valid pulse positions (p_n) and associated tracks ($\mathcal{T}_t$)
$0, 5$	$0, 5, 10, 15, 20, 25, 30, 35$ ($\mathcal{T}_0$)
$1, 6$	$1, 6, 11, 16, 21, 26, 31, 36$ ($\mathcal{T}_1$)
$2, 7$	$2, 7, 12, 17, 22, 27, 32, 37$ ($\mathcal{T}_2$)
$3, 8$	$3, 8, 13, 18, 23, 28, 33, 38$ ($\mathcal{T}_3$)
$4, 9$	$4, 9, 14, 19, 24, 29, 34, 39$ ($\mathcal{T}_4$)

5.5.2 3GPP EFR (ACELP)

As a second application example, ACELP JSCDH shall be applied to the 3GPP EFR codec [ETSI 1998, Järvinen et al. 1997] which, as G.729, also uses 5 ms subframes, a sampling rate of $f_\mathrm{s} = 8\,\mathrm{kHz}$, and a fixed codebook search based on a depth-first tree search approach with preselection of some pulses.

In total, $N_\mathrm{P} = 10$ pulses are placed in $N_\mathrm{T} = 5$ tracks with two pulses for each track, see Table 5.3. Also here, pulse signs $s_\mathrm{p}(p_n)$ are predetermined out of the loop (Section 5.4.3), but, as two pulses are coded for one track, an efficient encoding scheme is employed for the pulse signs which makes the bitstream order of the pulse position indices ambiguous. The size of the full search tree with its $1.07 \cdot 10^9$ leafs and $1.23 \cdot 10^9$ nodes is immense and a strong complexity reduction is inevitable for any practically viable codebook search algorithm.

In the following, the standardized codebook search algorithm based on a depth-first tree search with preselection is summarized. Then, a new codebook partitioning that is suited to hide $R_\mathrm{DH} = 10\,\mathrm{bit}/5\,\mathrm{ms} = 2\,\mathrm{kbit/s}$ of data in the pulse position indices is introduced along with the respective steganographic codebook search algorithm. Other configurations (with hidden bit rates down to 200 bit/s) are summarized in Appendix C.

Standard Codebook Search

The 3GPP EFR codec reduces the full FCB search tree by a depth-first approach and by fixing two pulses at the global or local maximum magnitude of the likelihood vector $\mathbf{b} = (b(0), b(1), \ldots, b(L_\mathrm{SF} - 1))^\mathrm{T}$. The standard defines this vector as the sum of the normalized long-term prediction residual $\mathbf{u}_\mathrm{LTP}$ (see [ETSI 1998]) and the normalized vector $\mathbf{d}$ from (5.16):

$$\mathbf{b} = \frac{\mathbf{u}_\mathrm{LTP}}{||\mathbf{u}_\mathrm{LTP}||} + \frac{\mathbf{d}}{||\mathbf{d}||}. \tag{5.35}$$

In the FCB search, the first pulse is fixed on the *global* maximum of $|b(k)|$. The track that is associated with the position of this maximum is denoted by $\mathcal{T}_{t_0}$. Then, four iterations are carried out. In these iterations, a second pulse is tentatively set

Table 5.4: Restricted FCB structure for 3GPP EFR, $R_{\mathrm{DH}} = 2\,\mathrm{kbit/s}$.

pulse number (n)	first valid position	second valid position	track
5	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_0) \oplus m_0\big)$	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_0) \oplus (m_0+4)\big)$	$\mathcal{T}_0^{m_0}$
6	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_1) \oplus m_1\big) + 1$	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_1) \oplus (m_1+4)\big) + 1$	$\mathcal{T}_1^{m_1}$
7	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_2) \oplus m_2\big) + 2$	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_2) \oplus (m_2+4)\big) + 2$	$\mathcal{T}_2^{m_2}$
8	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_3) \oplus m_3\big) + 3$	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_3) \oplus (m_3+4)\big) + 3$	$\mathcal{T}_3^{m_3}$
9	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_4) \oplus m_4\big) + 4$	$5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_4) \oplus (m_4+4)\big) + 4$	$\mathcal{T}_4^{m_4}$

on the *local* maximum of $|b(k)|$ within each of the four tracks that have not yet been occupied by the first pulse. Thereby, the employed track order is defined by the track indices $t_\nu = [(t_0 + \nu) \bmod N_{\mathrm{T}}]$ for $\nu \in \{1, \ldots, N_{\mathrm{T}} - 1\}$. The remaining eight pulses are then optimized in pairs.

Assuming a fixed first pulse in track $\mathcal{T}_{t_0}$, the following scheduling is used in the search algorithm to optimize the remaining nine pulses, the braces denoting pulse tracks that are optimized jointly:

1. $\mathcal{T}_{t_1}$, $(\mathcal{T}_{t_2}, \mathcal{T}_{t_3})$, $(\mathcal{T}_{t_4}, \mathcal{T}_{t_0})$, $(\mathcal{T}_{t_1}, \mathcal{T}_{t_2})$, $(\mathcal{T}_{t_3}, \mathcal{T}_{t_4})$
2. $\mathcal{T}_{t_2}$, $(\mathcal{T}_{t_3}, \mathcal{T}_{t_4})$, $(\mathcal{T}_{t_0}, \mathcal{T}_{t_1})$, $(\mathcal{T}_{t_2}, \mathcal{T}_{t_3})$, $(\mathcal{T}_{t_4}, \mathcal{T}_{t_1})$
3. $\mathcal{T}_{t_3}$, $(\mathcal{T}_{t_4}, \mathcal{T}_{t_0})$, $(\mathcal{T}_{t_1}, \mathcal{T}_{t_2})$, $(\mathcal{T}_{t_3}, \mathcal{T}_{t_4})$, $(\mathcal{T}_{t_1}, \mathcal{T}_{t_2})$
4. $\mathcal{T}_{t_4}$, $(\mathcal{T}_{t_0}, \mathcal{T}_{t_1})$, $(\mathcal{T}_{t_2}, \mathcal{T}_{t_3})$, $(\mathcal{T}_{t_4}, \mathcal{T}_{t_1})$, $(\mathcal{T}_{t_2}, \mathcal{T}_{t_3})$

The scheduling for the latter three iterations is derived from the first one by a cyclical shift of the track assignment for the last nine pulses. The corresponding search tree has $N_{\mathrm{L}} = 1008$ leafs and $N_{\mathrm{N}} = 1157$ nodes, thus facilitating an implementation with manageable complexity.

Codebook Partitioning for Steganography

The proposed codebook partitioning for the 3GPP EFR codec is based on the pulse position *indices* i_n according to (5.25) instead of the actual positions p_n. The procedure follows the description of Section 5.4.4.

To solve the parity condition (5.26) for the admissible indices i_{n_2}, the first index i_{n_1} must already be known. Therefore, only the positions for the *second* pulse in each track (i.e., $p_5, p_6, \ldots, p_9$) are restricted here. To achieve the desired bit rate of 2 kbit/s (or 10 bits per 5 ms subframe), $N_{\mathrm{T}} = 5$ (sub-)messages $m_j \in \{0, \ldots, M-1\}$ with $j \in \{0, \ldots, N_{\mathrm{T}} - 1\}$ and $\mathrm{ld}\, M = 2$ bits each must be

Table 5.5: Gray index assignment for pulse pos. indices in 3GPP EFR.

i	0	1	2	3	4	5	6	7
$i_{\text{bin.}}$	000	001	010	011	100	101	110	111
$\mathcal{G}(i)$	0	1	3	2	6	4	5	7
$[\mathcal{G}(i)]_{\text{bin.}}$	000	001	011	010	110	100	101	111

embedded within each of the 3-bit indices $i_5, i_6, \ldots, i_9$. The total rate of 2 kbit/s results from (5.29). Hence, $2^3/2^2 = 2$ admissible pulse positions remain for each of the pulses $p_5, p_6, \ldots, p_9$. According to (5.28), the restricted track sets are:

$$\mathcal{T}_t^{m_j} = \{N_\mathrm{T} \cdot (i_{n_1} \oplus m_j) + t,\ N_\mathrm{T} \cdot (i_{n_1} \oplus (m_j + M)) + t\}\,. \tag{5.36}$$

with $t \in \{0, 1, \ldots, N_\mathrm{T} - 1\}$ and $M = 4$. To obtain a *concrete* codebook partitioning, suitable values for n_1 and for the (sub-)message indices j must be selected. To properly resolve the bit position ambiguity of i_t and i_{t+N_T} in the bitstream, n_1 must be equal to t so that i_t and i_{t+N_T} are algebraically coupled as defined in (5.26). For the (sub-)message indices j, it is convenient to let $j = t$ as well. These choices result in the final sets

$$\mathcal{T}_t^{m_t} = \{N_\mathrm{T} \cdot (i_t \oplus m_t) + t,\ N_\mathrm{T} \cdot (i_t \oplus (m_t + M)) + t\}\,. \tag{5.37}$$

These restricted track sets can now be used to optimize the pulse positions $p_5, p_6, \ldots, p_9$ and the corresponding indices $i_5, i_6, \ldots, i_9$.

As an extension of (5.37), it is beneficial (as shown below) to take the standardized *Gray code index assignment* $\mathcal{G}(i_n)$ [ETSI 1998] of the pulse position codewords i_n into account. The employed Gray code is detailed in Table 5.5. To incorporate $\mathcal{G}(\cdot)$ into the codebook partitioning, Equation (5.26) is modified so that, effectively, the parity condition is enforced in the Gray coded domain:

$$[\mathcal{G}(i_{n_1}) \oplus \mathcal{G}(i_{n_2})] \mod M \stackrel{!}{=} m_j. \tag{5.38}$$

Solving (5.38) for i_{n_2} yields $L_\mathrm{SF}/(N_\mathrm{T} \cdot M) = 40/(5 \cdot 4) = 2$ solutions, cf. (5.27):

$$i_{n_2} = \mathcal{G}^{-1}\big(\mathcal{G}(i_{n_1}) \oplus (m_j + \nu \cdot M)\big) \quad \text{with} \quad \nu \in \{0, 1\}\,. \tag{5.39}$$

Using (5.25) and setting $n_1 = j = t$, the restricted track sets finally become:

$$\mathcal{T}_t^{m_t} = \big\{N_\mathrm{T} \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus m_t\big) + t,\ N_\mathrm{T} \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus (m_t + M)\big) + t\big\}\,. \tag{5.40}$$

The complete set of admissible pulse positions for $p_5, \ldots, p_9$ is given in Table 5.4 (the track sets for $p_0, \ldots, p_4$ are unchanged from Table 5.3).

Naturally, $\mathcal{G}(\cdot)$ and $\mathcal{G}^{-1}(\cdot)$ could also be omitted to achieve a valid codebook partitioning, see (5.37), but the reason for their inclusion is an increased robustness of the hidden information to transmission errors. The Gray coding has the undesired property that a single bit error in a Gray coded codeword $\mathcal{G}(i_n)$ may result in *two* (or even *three*) bit errors within the decoded codeword i_n.[8] Hence, the error rate for the steganographic data would be needlessly increased by a significant amount. Further note that, when decoding the hidden bits, the Gray coded versions of the pulse position codewords are directly available from the codec bitstream, i.e., no further decoding operation is necessary. The decoding equation is essentially identical to (5.38), concretely:

$$\hat{m}_t = \left[\underbrace{\mathcal{G}\left(i_t\right)}_{\substack{\text{directly available} \\ \text{from the bitstream}}} \oplus \underbrace{\mathcal{G}\left(i_{t+5}\right)}_{\substack{\text{directly available} \\ \text{from the bitstream}}} \right] \bmod M \quad \text{for} \quad t \in \{0, \ldots, N_\mathrm{T} - 1\}\,. \tag{5.41}$$

Steganographic Codebook Search

Using the restricted FCB of Table 5.4 in conjunction with the *standard* search procedure that has been described above would result in a significantly degraded speech quality, because instead of 1008 leafs and 1157 nodes, now merely 336 leafs and 413 nodes would be examined. An extension of the search space is therefore very important.

In the steganographic codebook search that has been described in Section 5.5.1 for the ITU-T G.729A codec, the speech quality could be maximized by employing a *full* tree search that was only restricted by the steganographic index constraints. However, in the case of the 3GPP EFR codec, this approach would still be too complex. The full tree search examines $1.07 \cdot 10^9$ candidate codevectors. With the above codebook partitioning to achieve a hidden rate of 10 bits per subframe, 2^{10} sub-codebooks have been established. Still, $1.07 \cdot 10^9 / 2^{10} \approx 1.04 \cdot 10^6$ candidate vectors remain for each sub-codebook. Therefore, the goal here shall be to achieve a complexity that is not much higher than that of the standardized codebook search algorithm.

As one possible solution, a modified depth-first search with pulse regrouping is used, cf. Section 5.4.5. The algorithm retains the fixed position of the first pulse on the maximum magnitude of $\mathbf{b}$ and also the four iterations, where the second pulse is tentatively placed on the local maxima within the unoccupied tracks. The search space expansion is actually achieved by jointly optimizing the positions of *four* instead of two pulses.

[8] If for example the first bit of the Gray code 110 is disturbed, i.e., the codeword $\underline{0}10$ is received, i_n is changed in *all three* bit positions (from 100 to 011).

Given a choice for first pulse from track $\mathcal{T}_{t_0}$ (max. of $|b(k)|$), the concrete scheduling for the steganographic codebook search can be stated as:

1. $\mathcal{T}_{t_1}$, $(\mathcal{T}_{t_0}^{m_0}, \mathcal{T}_{t_1}^{m_1}, \mathcal{T}_{t_2}, \mathcal{T}_{t_3})$, $(\mathcal{T}_{t_2}^{m_2}, \mathcal{T}_{t_3}^{m_3}, \mathcal{T}_{t_4}, \mathcal{T}_{t_4}^{m_4})$
2. $\mathcal{T}_{t_2}$, $(\mathcal{T}_{t_0}^{m_0}, \mathcal{T}_{t_2}^{m_2}, \mathcal{T}_{t_3}, \mathcal{T}_{t_4})$, $(\mathcal{T}_{t_3}^{m_3}, \mathcal{T}_{t_4}^{m_4}, \mathcal{T}_{t_1}, \mathcal{T}_{t_1}^{m_1})$
3. $\mathcal{T}_{t_3}$, $(\mathcal{T}_{t_0}^{m_0}, \mathcal{T}_{t_3}^{m_3}, \mathcal{T}_{t_4}, \mathcal{T}_{t_1})$, $(\mathcal{T}_{t_4}^{m_4}, \mathcal{T}_{t_1}^{m_1}, \mathcal{T}_{t_2}, \mathcal{T}_{t_2}^{m_2})$
4. $\mathcal{T}_{t_4}$, $(\mathcal{T}_{t_0}^{m_0}, \mathcal{T}_{t_4}^{m_4}, \mathcal{T}_{t_1}, \mathcal{T}_{t_2})$, $(\mathcal{T}_{t_1}^{m_1}, \mathcal{T}_{t_2}^{m_2}, \mathcal{T}_{t_3}, \mathcal{T}_{t_3}^{m_3})$,

where, again, $t_\nu = [(t_0 + \nu) \bmod N_\mathrm{T}]$ for $\nu \in \{1, \ldots, N_\mathrm{T} - 1\}$. The number of tree leafs amounts to $N_\mathrm{L} = 1272$ and the number of visited nodes is $N_\mathrm{N} = 1589$. These figures are actually relatively close to the original numbers (1008 and 1157). Therefore, a similar speech quality as with the standard codebook search algorithm can be expected. Note that, in the above track scheduling, the restricted tracks have been scheduled as early as possible which considerably reduces the number of visited tree nodes. An unfavorable scheduling where the restricted tracks would be scheduled as late as possible (while preserving the pulse grouping in quadruples) would lead to 2309 tree nodes.

Finally, the entire steganographic codebook search algorithm to hide 2 kbit/s in the EFR bitstream, as implemented in software, is summarized in the box below. In addition to the present algorithm for $R_\mathrm{DH} = 2\,\mathrm{kbit/s}$, Appendix C discusses several other data hiding modes for 3GPP EFR that offer lower hidden data rates (down to 200 bit/s).

1. Let $(t_0, \ldots, t_{N_\mathrm{T}-1})$ with $t_{\nu+1} = [(t_\nu + 1) \mod N_\mathrm{T}]$ and $N_\mathrm{T} = 5$ denote the preselected permutation of track indices which ensures that the maximum of $|b(k)|$, see (5.35), lies in track $\mathcal{T}_{t_0}$, i.e., $t_0 = [\arg\max_k |b(k)|] \mod N_\mathrm{T}$.

2. Fix the first pulse p_{t_0} on the global maximum of $|b(k)|$, i.e., $p_{t_0} = \arg\max_{k \in \mathcal{T}_{t_0}} |b(k)|$.

3. Compute the admissible values for the second pulse $p_{t_0+\mathrm{N}_t}$, i.e., establish the restricted track set $\mathcal{T}_{t_0}^{m_{t_0}}$ according to Table 5.4. Equation (5.25) can be used to obtain the pulse position index i_{t_0}.

4. Initialize the iteration counter.

5. Set the pulse p_{t_1} to the position of the maximum of $|b(k)|$ within track $\mathcal{T}_{t_1}$, i.e., $p_{t_1} = \arg\max_{k \in \mathcal{T}_{t_1}} |b(k)|$.

6. Compute the admissible values for $p_{t_1+N_\mathrm{T}}$, i.e., establish the restricted track set $\mathcal{T}_{t_1}^{m_{t_1}}$ according to Table 5.4.

7. Jointly optimize $p_{t_0+N_\mathrm{T}} \in \mathcal{T}_{t_0}^{m_{t_0}}$, $p_{t_1+N_\mathrm{T}} \in \mathcal{T}_{t_1}^{m_{t_1}}$, $p_{t_2} \in \mathcal{T}_{t_2}$, and $p_{t_3} \in \mathcal{T}_{t_3}$. The partial CELP criterion $C_{\mathcal{P}_I}^2 / E_{\mathcal{P}_I}$, as defined by (5.20) and (5.21), is maximized. The set $\mathcal{P}_I$ comprises the pulse positions that are known so far.

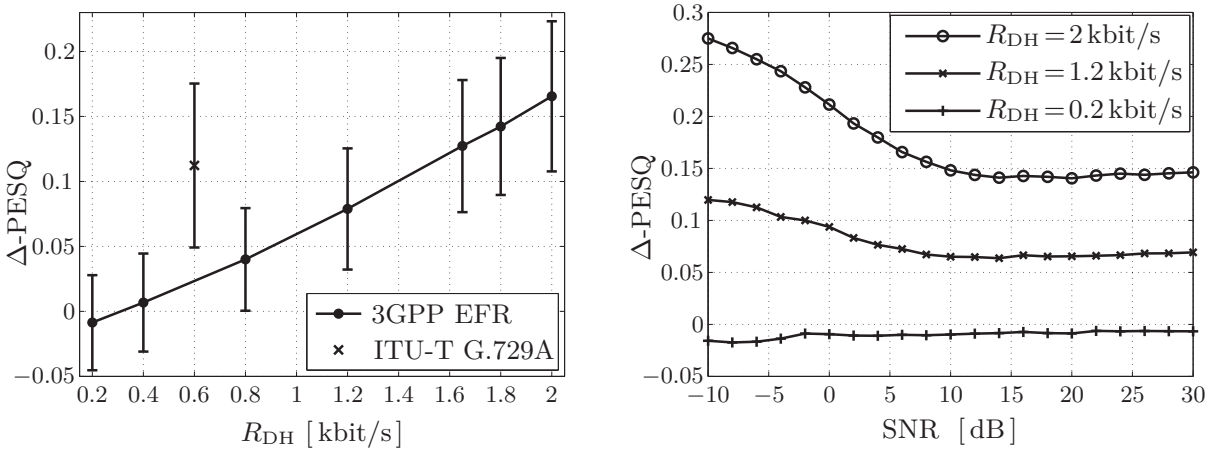
8. Compute the admissible values for $p_{t_2+N_\mathrm{T}}$ and $p_{t_3+N_\mathrm{T}}$, i.e., establish the restricted track sets $\mathcal{T}_{t_2}^{m_{t_2}}$ and $\mathcal{T}_{t_3}^{m_{t_3}}$ according to Table 5.4.

9. Jointly optimize $p_{t_2+N_\mathrm{T}} \in \mathcal{T}_{t_2}^{m_{t_2}}$, $p_{t_3+N_\mathrm{T}} \in \mathcal{T}_{t_3}^{m_{t_3}}$, $p_{t_4} \in \mathcal{T}_{t_4}$, and $p_{t_4+N_\mathrm{T}} \in \mathcal{T}_{t_4}^{m_{t_4}}$. The CELP criterion $C_{\mathcal{P}_{II}}^2 / E_{\mathcal{P}_{II}}$ is maximized whereby the set $\mathcal{P}_{II}$ contains all N_P pulses. In the course of the optimization, the admissible values for $p_{t_4+N_\mathrm{T}}$ must also be updated suitably (recomputation of $\mathcal{T}_{t_4}^{m_{t_4}}$ for each new value of p_{t_4}).

10. Compare the new candidate codevector with the best vector of all previous iterations. Remember the selected codevector, if an improvement is obtained.

11. Cyclically shift the current permutation of track indices:
$t_\nu \leftarrow t_{(\nu \bmod (N_\mathrm{T}-1))+1}$ for $\nu \in \{1, \ldots, N_\mathrm{T} - 1\}$.

12. Increase the iteration counter and stop if four iterations have been carried out; otherwise go to Step 5.



(a) Clean speech ($\mathrm{PESQ}_{\mathrm{std.,EFR}} = 4.12$, $\mathrm{PESQ}_{\mathrm{std.,G.729}} = 3.74$)

(b) Noisy speech (Codec: 3GPP EFR, $\mathrm{PESQ}_{\mathrm{std.,EFR}} = 4.03 \ldots 4.25$)

Figure 5.11: Results for JSCDH in ACELP codecs. Evaluation of average PESQ loss compared to standard codec (Δ-PESQ).

5.5.3 Evaluation and Test Results

In the following, the proposed data hiding algorithms shall be evaluated using both objective and subjective test methods. Of particular interest is the potential loss in narrowband speech quality due to the hidden data. For the present evaluation, the hidden messages are in general generated randomly.

Objective Evaluation (PESQ)

The ITU-T PESQ tool [ITU-T 2001, Rix et al. 2001] is suited to evaluate the quality of processed (decoded) narrowband speech signals based on the respective reference (original) samples. The PESQ score ranges from approx. 1.0 (very bad quality) to 4.6 (excellent quality) and is assumed to approximate the outcome of a subjective listening test based on the classical MOS (mean opinion score) scale. In the context of JSCDH in speech codecs, the PESQ *loss* is important. Therefore, the PESQ scores for both the standard codecs ($\mathrm{PESQ}_{\mathrm{std.}}$) and for the modified steganographic codecs ($\mathrm{PESQ}_{\mathrm{steg.}}$) are measured and averaged over the entire NTT speech corpus [NTT 1994] at an "active speech level" according to [ITU-T 1993*b*] of −26 dBov. The impact of the data hiding operation is then assessed via the differential score

$$\Delta\text{-PESQ} \doteq \mathrm{PESQ}_{\mathrm{std.}} - \mathrm{PESQ}_{\mathrm{steg.}}. \tag{5.42}$$

The results (avg. ± std. dev.) for JSCDH in the ITU-T G.729A codec (Section 5.5.1) and the 3GPP EFR codec (Section 5.5.2) under clean speech conditions are shown in Figure 5.11(a). The additional hidden bit rate modes for the EFR as shown in the figure are summarized in Appendix C. In general, values of $\Delta\text{-PESQ} < 0.2$ indicate that the quality loss is very moderate or even negligible. In fact, at a low hidden bit rate of $R_{\mathrm{DH}} = 200$ bit/s for the EFR codec, even a small quality *improvement* over the standard codec can be observed. This can be attributed to the fact that the implemented steganographic codebook search

algorithm for $R_{\mathrm{DH}} = 200\,\mathrm{bit/s}$ is relatively complex and much more candidate codevectors are examined than in the standard. Some results for noisy speech conditions, using the "m109" noise of the Noisex-92 database [*Noisex-92: Database of recording of various noises* 1992], are shown in Figure 5.11(b). The noise obviously has no or little impact, at least for low and medium hidden bit rates. For high hidden bit rates ($R_{\mathrm{DH}} = 2\,\mathrm{kbit/s}$), the excellent performance under clean speech can be maintained down to SNR values of ca. 5 dB. Below this value, the fixed codebook contribution apparently gains too much importance, resulting in a somewhat higher (but nevertheless acceptable) quality loss.

The abscissa of Figure 5.11(a) shows the *absolute* hidden data rate R_{DH}. To provide a more fair comparison among different methods and codecs, Figure 5.12 displays the same results plotted over the *relative* hidden rate, i.e., R_{DH} divided by the codec bit rate R_{codec}. As additional references, several values of Δ-PESQ for other data hiding methods that have been published in the literature (cf. Section 5.3.3) are included. The direct comparison confirms an exceptionally low speech quality loss of the proposed JSCDH method over the entire range of bit rates. Still, in Figure 5.12, it is obvious that the proposed data hiding method in the G.729A codec performs somewhat worse when compared to the EFR solution. For example, EFR data hiding with a bit rate of $R_{\mathrm{DH}} = 1.65\,\mathrm{kbit/s}$ (relative hidden rate: 13.5%) leads to the same quality loss as G.729A data hiding with $R_{\mathrm{DH}} = 0.6\,\mathrm{kbit/s}$ (relative hidden rate: 7.5%). A possible explanation for this behavior lies in the simple structure of the G.729 fixed codebook which only allows *one* pulse per track (cf. Table 5.1). When the restricted codebook according to Table 5.2 is used, some pulse positions within the excitation vector $\mathbf{c}$ cannot be occupied anymore. The EFR coder, in contrast, places *two* pulses in one track (cf. Table 5.3) and, according to Table 5.4, only *one* of these two pulses is restricted for data hiding. The first pulse in each track may assume *all* possible positions which leaves sufficient degrees of freedom to constrain the speech quality loss.

Objective Evaluation (HWR)

As a second objective quality measure, the "Host-to-Watermark" ratio (HWR) as defined in Section 5.1.1 is exemplarily evaluated for data hiding with 2 kbit/s in the 3GPP EFR codec. Here, the HWR is the logarithmic ratio between the power of the unmodified decoded speech and the power of the "watermark signal" which is defined as the difference signal between the unmodified decoded speech and the decoded speech with hidden data. The HWR for the example waveform shown in Figure 5.13 is ca. 20.3 dB. Measurements based on a larger speech database reveal that the steganographic EFR codec with its 2 kbit/s of hidden data exhibits an averaged HWR of 19.3 dB. To provide a meaningful comparison, also the average SNR that is incurred when migrating from the standard *floating point* implementation to the standard *fixed point* implementation of this codec has been measured. The respective measurement yields 20.3 dB which is merely 1 dB above the HWR of the data hiding scheme.

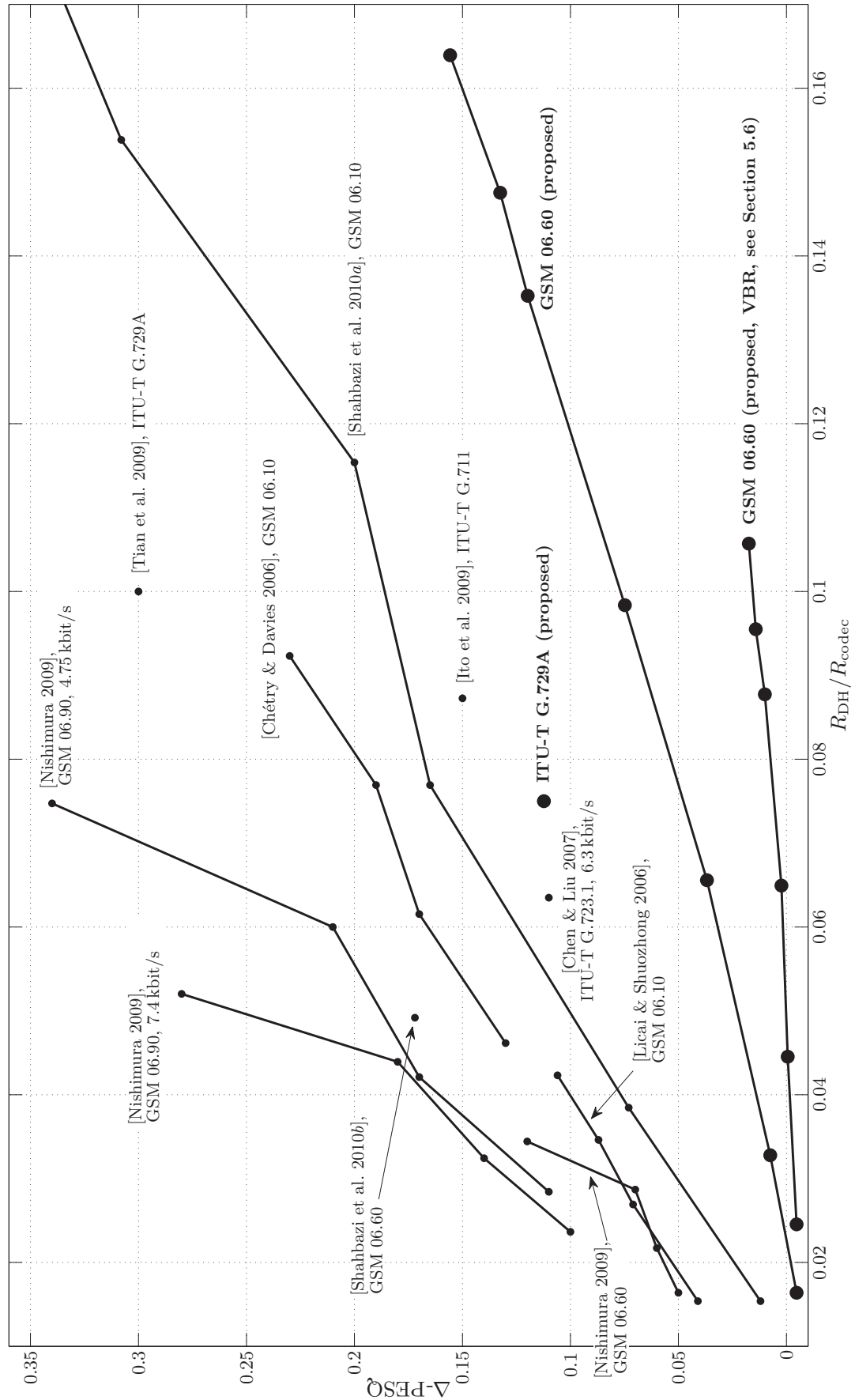


Figure 5.12: Comparison of speech codec data hiding algorithms based on a relative hidden bit rate ($R_{\mathrm{DH}}/R_{\mathrm{codec}}$). Evaluation of average PESQ loss compared to standard codec (Δ-PESQ). Values are taken from the respective publications. (VBR: Variable Bit Rate)

Subjective Listening Test (ABX)

In order to assess if any difference can be *perceived* between decoded speech samples from both versions of the EFR codec, an ABX listening test was conducted where eleven experienced listeners—using closed back headphones in a quiet environment—had to decide whether the presented test sample X was equal to reference A or B. The options A and B have been randomly assigned to "standard EFR coded speech" and "speech with 2 kbit/s of hidden data." The data hiding mode with 2 kbit/s has been chosen as a worst case scenario since the previous PESQ evaluation indicated the highest potential quality impact in this case. For the test, six short utterances from the NTT corpus [NTT 1994] (three female and three male speakers) have been processed by both versions of the coder and presented to the subjects. Each utterance had to be judged four times. Before making a judgment, the samples A, B, and X could be played ad libitum. In total, $11 \cdot 4 \cdot 6 = 264$ votes have been received, and only in 162 cases (61%), the correct decision was made. A *statistically significant* number of correct votes was only observed for the female speech samples (66%, 66%, 70%). All listeners agreed that the (possibly) perceived differences were very hard to detect and that the difference between samples A and B in terms of speech quality is very small.

Example Waveform and "Segmental PESQ Scores"

To gain a more detailed insight into the characteristics of the proposed JSCDH scheme, a concrete example waveform is presented in the upper graph of Figure 5.13. The 3GPP EFR codec with 2 kbit/s of hidden data has been used to produce this example. In the graph, apart from the output of the steganographic codec, also the *difference signal* between the outputs of the unmodified (standard) EFR codec and the steganographic codec version is plotted.

The impact of the data hiding operation is obviously dependent on the characteristics of the current signal segment, i.e., speech sound. While the difference signal is relatively small for voiced speech (e.g., around time index 1.0 s), it may nearly attain the level of unvoiced speech (e.g., around time index 0.6 s). This observation is easily explained when recalling that the data is hidden in the index for the fixed (stochastic) codebook of the ACELP codec which is particularly important in regenerating the typical noise characteristics of unvoiced speech. In voiced speech, the fixed codebook is less important while the adaptive codebook provides the major contribution to the final excitation signal.

To analyze the *perceptual* impact of the data hiding operation on different speech segments, a "segmental" PESQ score difference has been computed. Therefore, the ITU-T PESQ tool [ITU-T 2001, Rix et al. 2001] for objective judgment of the speech quality of narrowband signals has been modified to map its internal framewise signal representation (16 ms frames) directly to the usual PESQ scale (MOS-LQO, 1.0 – 4.6) without any further temporal processing and averaging so

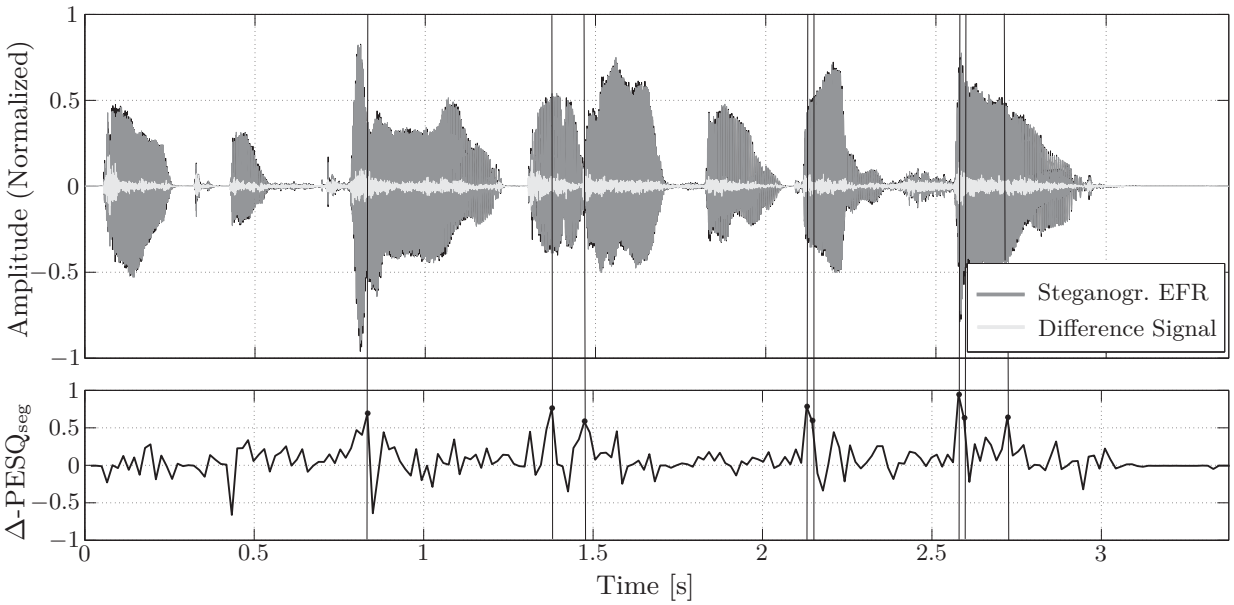


Figure 5.13: Example waveforms for data hiding with 2 kbit/s in the 3GPP EFR codec. Spoken text of the female speaker: "*Oak is strong and also gives shade.*" An objective quality evaluation is given in terms of a segmental PESQ score difference (Δ-PESQ$_{\text{seg}}$).

that detrimental effects on speech quality can be analyzed with a much finer temporal granularity. The lower graph in Figure 5.13 shows the difference in segmental PESQ scores between the output of the standard EFR codec and the output of the steganographic EFR codec. Positive values of Δ-PESQ$_{\text{seg}}$ indicate that the standard version of the codec is (momentarily) better than the steganographic version and vice versa for negative values. It can be observed that, although the difference signal is rather large in unvoiced speech, this has no significant impact on speech quality which can be explained by the fact that the actual realization of the noise process is not important for human auditory perception. However, there are a few speech segments where the data hiding actually leads to significantly lower segmental PESQ scores. In the figure, time indices where Δ-PESQ$_{\text{seg}} \geq 0.5$ are marked with vertical black lines. Thereby, as a tendency, it can be observed that *onset* segments (especially voiced onsets) and segments where a sudden *pitch change* occurs are particularly sensitive, an observation which is also confirmed by inspecting other speech samples than the presented one. The explanation here is that the fixed codebook contribution is essential in building up the vibration at the beginning of a voiced speech segment, because the adaptive codebook is not yet effective at this point (the memory of the adaptive codebook is still filled with unvoiced speech, noise, or silence). On the other hand, there are a few other segments where the quality has even improved over the standard codec. This is possible because the steganographic codebook search algorithm covers different parts of the algebraic codebook than the, also non-exhaustive, standard algorithm.

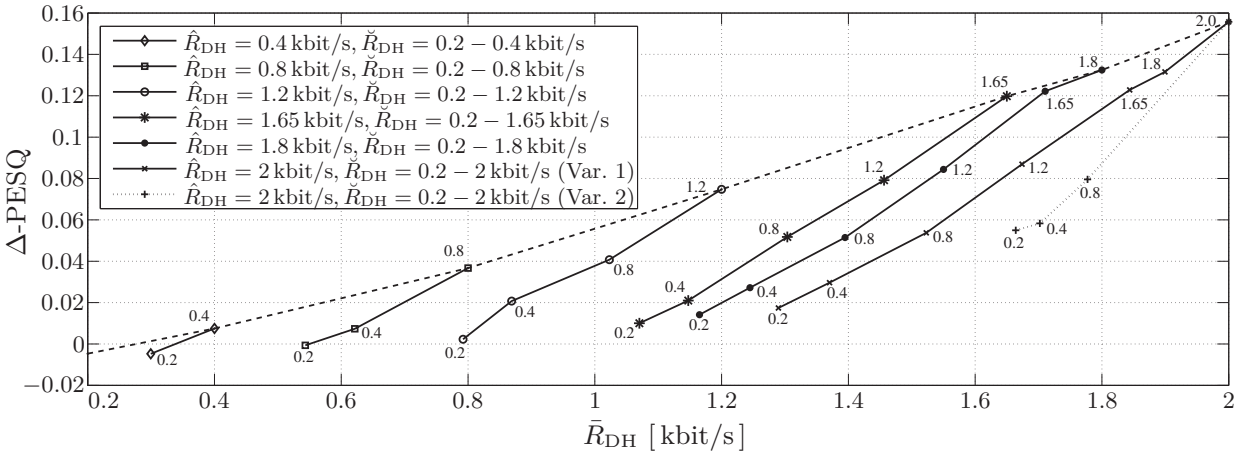


Figure 5.14: Results for JSCDH with variable rate in the 3GPP EFR codec. Evaluation of average PESQ loss compared to standard codec (Δ-PESQ). The high hidden bit rate $\hat{R}_{\mathrm{DH}}$ of each data hiding mode is indicated by the individual graph markers. The associated low hidden bit rate $\breve{R}_{\mathrm{DH}}$ for critical speech segments is directly annotated at the corresponding data points.

5.6 ACELP Data Hiding with Variable Bit Rate

The different impact of the data hiding procedure on specific speech segments which has been shown above can actually be "evened out" by adapting the hidden data rate (and therefore the degrees of freedom for the fixed codebook) to the characteristics of the current speech segment, leading to a data hiding scheme with *variable bit rate*. Here, variable rate data hiding shall be achieved with a combination of *two* hidden bit rate modes. Thereby, the lower hidden rate $\breve{R}_{\mathrm{DH}}$ shall be used for particularly quality-sensitive speech (sub)frames (as marked with the vertical lines in Figure 5.13) while the higher hidden rate $\hat{R}_{\mathrm{DH}}$ shall be used for the remaining (uncritical) speech (sub)frames.

The switching between the two bit rate modes shall be controlled by a frame classifier. However, to be able to recover the hidden message, it is important that the outcome of the frame classification is available at the decoder side as well. If no hidden bits shall be sacrificed to embed the classification result itself into the codec bitstream, the classification must rely on quantized codec parameters that are available to both the encoder and the decoder. As an additional requirement, suitable quantized parameters must be available *before* the steganographic codebook search is conducted. For the 3GPP EFR codec, according to [ETSI 1998], the spectral envelope, the pitch lag, and the pitch gain parameters of the current speech frame (and in principle all parameters of previous frames) are usable to decide on the hidden bit rate mode ($\breve{R}_{\mathrm{DH}}$ or $\hat{R}_{\mathrm{DH}}$). It is important to note that the gain for the fixed codebook is *not* use-able for this classification task since it depends on the fixed codebook index. The fixed codebook index, in turn, depends on the chosen hidden bit rate.

As an example, a simple tree-based classificator [Breiman et al. 1984] has been trained based on the *current* quantized parameters (spectral envelope, pitch lag and gain). The frames in the training set have been labeled as *critical* or *uncritical* based on a certain threshold for the segmental PESQ measure in the respective frame. Using the trained classifier to switch between the two bit rates, the resulting speech quality loss (Δ-PESQ) has been evaluated for various combinations of $\breve{R}_{\mathrm{DH}}$ and $\hat{R}_{\mathrm{DH}}$. The results are plotted over the *average* hidden rate $\bar{R}_{\mathrm{DH}}$ in Figure 5.14. The quality of the standard codec can obviously be maintained up to an average bit rate of $\bar{R}_{\mathrm{DH}} \approx 1.3\,\mathrm{kbit/s}$ (for $\breve{R}_{\mathrm{DH}} = 0.2\,\mathrm{kbit/s}$ and $\hat{R}_{\mathrm{DH}} = 2\,\mathrm{kbit/s}$). The incurred PESQ loss is then almost zero ($\Delta\text{-PESQ} < 0.02$).

For $\hat{R}_{\mathrm{DH}} = 2\,\mathrm{kbit/s}$, two different $\mathrm{PESQ}_{\mathrm{seg}}$-thresholds have been tested in the training. The results for the second variant are shown with the dotted line in the figure. In this configuration, much less speech frames are classified as "critical." Hence, higher average rates can be achieved, but a (very minor) quality loss of $\Delta\text{-PESQ} \approx 0.05$ cannot be avoided.

Relevance of the Δ-PESQ measure

Finally, some remarks shall be made concerning the observed range of values for the Δ-PESQ measure. At first sight, a PESQ difference of less than 0.1 might appear too small to draw dependable conclusions. However, the strict consistency which could be observed in *all* of the conducted experiments, e.g., Figure 5.11(a) and in particular Figure 5.14, suggests that the evaluation is generally valid. Naturally, as shown in the ABX test, even a value of $\Delta\text{-PESQ} \approx 0.16$ has a very small impact on the *subjective* listening impression.

5.7 Bandwidth Extension with Hidden Side Information

The steganographic transmission channel that becomes available with the data hiding techniques introduced above, shall now be used for bandwidth extension purposes. The corresponding transmission system, as a specialization of Figure 2.1, is shown in Figure 5.15.

In each signal frame (with index λ), the parameter set to describe the extension band signal $s_{\mathrm{eb}}(k)$ is encoded into a steganographic message $m(\lambda)$ which is hidden inside the codec bitstream. The marked bitstream is then transmitted, e.g., over a legacy transmission system involving GSM radio access. At the receiving end, a standard decoder produces the signal $\hat{s}_{\mathrm{bb}}(k)$. If the receiver is aware of the hidden information, the hidden message can be recovered as $\hat{m}(\lambda)$ and the corresponding bandwidth extension parameters and finally the artificial extension band signal $\hat{s}_{\mathrm{eb}}(k)$ can be obtained. The signal $\hat{s}_{\mathrm{eb}}(k)$ is then combined with the baseband signal to yield the bandwidth extended output $\hat{s}_{\mathrm{bwe}}(k')$.

In the following, a concrete implementation example of a steganographic wideband speech transmission system shall be examined more closely.

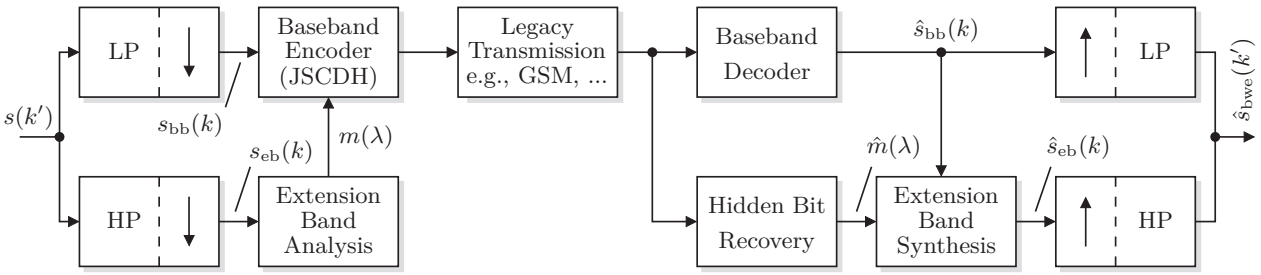


Figure 5.15: Backwards compatible transmission system using bandwidth extension with hidden side information.

5.7.1 The E^{2}FR Codec

In principle, all parameter sets and analysis/synthesis schemes as discussed in Chapters 2 and 3 can be used in a "steganographic wideband codec" if the available hidden bit rate is sufficient. As a practically relevant application example, the steganographic version of the 3GPP EFR codec (Section 5.5.2) has been enhanced with the TDBWE bandwidth extension module [Geiser et al. 2007a] (see Section 3.2), thus forming an "Enhanced EFR Codec" (E^{2}FR).

At the encoder, a half-band IIR QMF filterbank is used to split the wideband input signal $s(k')$ into a baseband signal $s_{\mathrm{bb}}(k)$ and an extension band signal $s_{\mathrm{eb}}(k)$. The corresponding synthesis filterbank combines $\hat{s}_{\mathrm{bb}}(k)$ and $\hat{s}(k)$ and applies a (partial) phase equalization, see [Löllmann et al. 2009] and Section 2.1.2. The TDBWE algorithm is used to encode and resynthesize the extension band signal. Originally, this algorithm has been developed as a part of this thesis for ITU-T Rec. G.729.1. However, an adaptation to the 3GPP EFR codec is easily accomplished as the only point of direct interaction between the baseband codec and the TDBWE algorithm is the high band excitation generation. In particular, the pitch lag parameter and the energies of the ACB and FCB contributions of the baseband CELP codec are required to generate the TDBWE excitation signal. This codec setup, without the steganographic components, is already discussed in [Jung et al. 2008]. In this thesis, the TDBWE bitstream with its bit rate of 1.65 kbit/s is included as steganographic payload in the EFR bitstream. Therefore, an ACELP JSCDH scheme with 1.65 kbit/s—which has been especially designed to match the TDBWE bit rate—is used, see Appendix C for details. The enhanced decoder can extract the steganographic information and synthesize $\hat{s}_{\mathrm{eb}}(k)$. A legacy decoder will ignore the hidden payload and output $\hat{s}_{\mathrm{bb}}(k)$.

The proposed E^{2}FR codec is included as one test condition (CuT-C) in the comparative quality test in Chapter 6. The respective results of the *subjective* DCR test (Section 6.2) and of the *objective* quality assessment (Appendix D) reveal that the proposed codec offers excellent wideband quality for its bit rate of 12.2 kbit/s. Thereby, the quality impact of the data hiding operation is shown to be negligible. Hence, full backwards compatibility with legacy systems which use the EFR codec can be guaranteed *without a noticeable quality loss*.

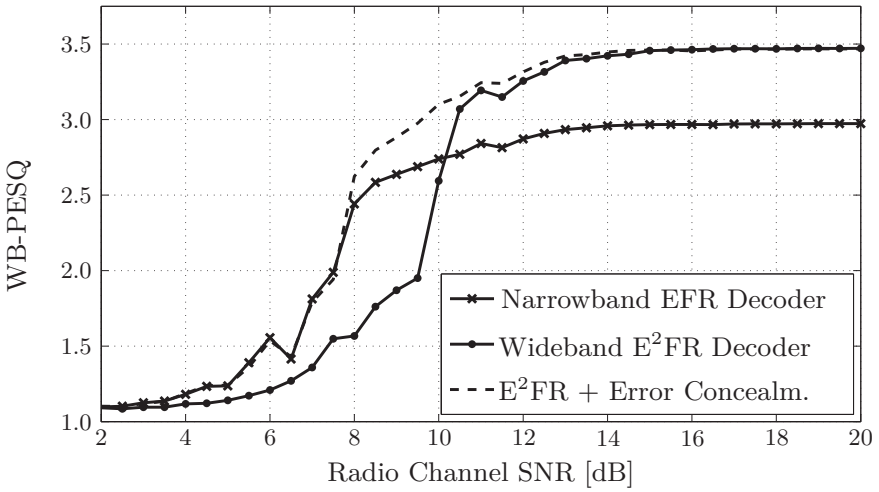


Figure 5.16: Results of the GSM link-level simulations for the TU50 channel profile: Objective *wideband speech quality* (WB-PESQ) for narrow-/wideband decoders, cf. [Geiser & Vary 2007*a*].

5.7.2 Transmission Over a GSM Radio Link

So far, an error-free transmission of the encoded bits and hence of the steganographic payload has been assumed. However, for practical systems, in particular when a radio link is involved, the question of robustness of the steganographic information becomes important. Therefore, it has to be considered that the hidden bits are embedded in the *least significant part* of the codec bitstream. In typical cellular networks, these bits are only weakly protected by the standardized channel coding schemes, e.g., [ETSI 2005]. On top of this, the entanglement of multiple codec bits to represent one hidden bit as, e.g., performed for data hiding in the 3GPP EFR codec (cf. Equation (5.38)), further increases the error probability for the hidden bits. For instance, if the ACELP codebook partitioning of Table 5.4 is used, a bit error in *either* of the indices i_t and i_{t+N_T} results in a bit error for the *same* hidden bit.

To assess the impact of bit errors due to radio transmission, link-level simulations for the "Enhanced Full Rate Traffic Channel" (TCH/EFS) of the GSM cellular communication system have been conducted. Concretely, the TU50 channel profile (typical urban scenario with a vehicle speed of 50 km/h) [ETSI 2001*a*] has been used and the speech quality has been measured with the wideband PESQ tool [ITU-T 2005, Takahashi et al. 2005] for a narrowband (standard) and for a wideband (steganographic) codec, see [Geiser & Vary 2007*a*]. In the experiment, EFR data hiding with a reduced hidden bit rate of $R_\mathrm{DH} = 400\,\mathrm{bit/s}$ was used and the bandwidth extension algorithm was adapted to this bit rate. Figure 5.16 shows the results of the speech quality assessment over the SNR of the radio channel. While the speech quality in the standard GSM system based on the narrowband EFR codec reaches an acceptable level at a radio channel SNR of 8 dB, the steganographic wideband codec needs a channel SNR of 10 dB to reach the same quality

level. For higher SNR values, there is a clear quality advantage.[9]

From these results it can be concluded that dedicated error detection and error concealment mechanisms should be used for the hidden bandwidth extension information. In a first feasibility study, the availability of a *perfect* "bad frame indicator" (BFI) for the hidden bits has been assumed. In case of bit errors within the current frame, an *estimated* high band signal (ABWE, see Chapter 4) has been inserted instead of the (erroneously) decoded version. As a result, the quality of the steganographic E^2FR wideband codec becomes in fact *consistently* better than the standard narrowband EFR over the *full* range of channel conditions (dashed curve in Figure 5.16). To finally obtain a *practical* solution for error concealment, it is advisable to slightly increase the hidden bit rate and to add a few CRC (Cyclic Redundancy Check) bits to the steganographic payload so that bit errors can be reliably detected. A hybrid approach, combining forward error correction and CRC bits, has been proposed in [Geiser, Mertz & Vary 2008].

5.8 Other Applications for Hidden Side Information

In principle, any kind of signaling and control information can be transmitted with the proposed ACELP JSCDH method without breaking backwards compatibility with legacy systems. Potential applications include *caller identification and verification* or even *secure communication* over the steganographic channel. In the latter case, the bitstream of a low bit rate codec such as the 1.2 kbit/s MELP coder [Wang et al. 2002] is directly used as steganographic payload. Aiming at telephony with better quality and enhanced features, the transmission of *spatial cues* can be interesting so that the decoded monaural speech can be binaurally rendered to a particular spatial location in a controlled manner [Geiser et al. 2011]. Another important use case for ACELP JSCDH in the field of high-quality speech transmission is *frame erasure concealment* in packet-switched communication systems. Typical CELP codecs are comparatively sensitive to frame erasures because the involved coding mechanisms are not memoryless. However, it could be shown that a limited amount of additional side information can significantly improve the performance of the concealment mechanism [Mertz 2011]. The steganographic channel inside the codec bitstream can then be used to transmit this information in a backwards compatible manner, see [Geiser, Mertz & Vary 2008].

Finally, the proposed JSCDH method can also be exploited to reduce the codec bit rate. The bit rate reduction is easily achieved by "transmitting" a *constant* steganographic message m and encoding the pulse positions with less bits, e.g., *one* instead of *three* bits for each restricted pulse of Table 5.4. In this respect, the data hiding procedure can be viewed as a *redundant encoding* of the admissible pulse position indices, cf. [Geiser & Vary 2008*b*].

[9] A *direct* comparison of narrowband *and* wideband conditions with the WB-PESQ tool may have limited significance. A crosscheck with subjective test results, e.g., Chapter 6, reveals that the quality gain of wideband speech is underestimated with the objective scores.

Chapter 6

Evaluation and Comparison

In this chapter, the different approaches for "High Definition Telephony" that have been discussed in this thesis, shall be evaluated and compared.

The obtained (wideband and super-wideband) *speech quality* was assessed with subjective listening tests. The test setup is described in Section 6.1 while the results are discussed in Sections 6.2 and 6.3, respectively. The super-wideband *audio quality* (e.g., music) is evaluated using an instrumental quality measure, see Section 6.4. For reference, the results of instrumental measurements of *speech quality* are provided in Appendix D.

6.1 Experimental Setup for the Subjective Listening Tests

The subjective listening tests were carried out according to the *degradation category rating* (DCR) method as standardized in [ITU-T 1996*c*]. In a DCR test, the subjects have to rate the *degradation* of a *processed* Sample B in relation to the *reference* Sample A. The *degradation rating* is then given on a five point scale ranging from "very annoying degradation" (1) to "inaudible degradation" (5). The mean score is referred to as DMOS (degradation mean opinion score).

However, in the present experiments, a *modified* DMOS scale has been employed since a rating of "very annoying degradation" was not expected considering the relatively benign distortions of the tested codecs. Instead, a better resolution of the scale at the upper end was desired so that smaller quality differences could be assessed. The five possible ratings of the modified DMOS scale (M-DMOS) are:

- degradation is inaudible (5)
- degradation is barely audible (4)
- degradation is clearly audible but not annoying (3)
- degradation is slightly annoying (2)
- degradation is annoying (1)

In the listening tests, Sample A (the reference) was a prefiltered [ITU-T 1995], but otherwise unprocessed (wideband or super-wideband) sample in the English language from the NTT corpus [NTT 1994] at an "active speech level" of

−26 dBov [ITU-T 1993*b*]. In total, twelve of these test samples (six male and six female talkers) have been encoded with the respective codecs and presented as Sample B. The material of these talkers has not been used for algorithm training.

The tests were conducted in a quiet environment (studio box) using Sennheiser HD600 open stereo headphones (diotic presentation). The headphones were driven by a dedicated amplifier with calibrated equalization. A comfortable presentation level was set by the subjects. Before each test session, the subjects were asked to listen to a demonstration sample that included examples of all processing variants. Eight subjects participated in each test, i.e., each test condition received $8 \times 12 = 96$ votes. Upon completion of a test session, ice cream was given away as gratification.

6.2 Wideband Speech Quality

Three scenarios are considered to extend narrowband signals (0.05 – 4 kHz) towards the wideband frequency range (0.05 – 7 kHz).

The 3GPP EFR codec (12.2 kbit/s) [ETSI 1998, Järvinen et al. 1997] is used in all of these proposal to encode the narrowband signal. The high band from 4 to 7 kHz is synthesized with the TDBWE algorithm that has originally been developed for ITU-T G.729.1, see Section 3.2 and [Geiser et al. 2007a] for details. The required adaptation of the algorithm to 3GPP EFR is easily accomplished which has also been shown in [Jung et al. 2008].

In the present codec proposals, the TDBWE parameter set is obtained in three different ways. The following "Codecs under Test" (CuT) are considered:

- **CuT-A: Embedded Coding** (Chapter 3)
  The TDBWE parameter set is determined at the encoder and *quantized* with the standardized bit rate of 1.65 kbit/s. This information is appended to the 12.2 kbit/s bitstream of the 3GPP EFR codec, thus forming an *embedded codec* with two bitstream layers. The sum bit rate is 13.85 kbit/s.

- **CuT-B: Receiver Based Parameter Estimation** (Chapter 4)
  Bayesian estimation with Hidden-Markov modeling is used to estimate the TDBWE parameter set based on features of the narrowband signal. No side information is transmitted in this case. The estimation algorithm has been configured with a 7-bit codebook and 16 Gaussian mixture components per state. The parameter set and the feature vector are defined in Section 4.3.6. CuT-B does not increase the bit rate.

- **CuT-C: Steganographic Parameter Transmission** (Chapter 5)
  Joint source coding and data hiding (JSCDH) with a hidden bit rate of 1.65 kbit/s is employed to transport the quantized TDBWE parameter set in a backwards compatible manner. The bit rate of CuT-C is still 12.2 kbit/s. The codec is identical with the E^2FR codec from Section 5.7.

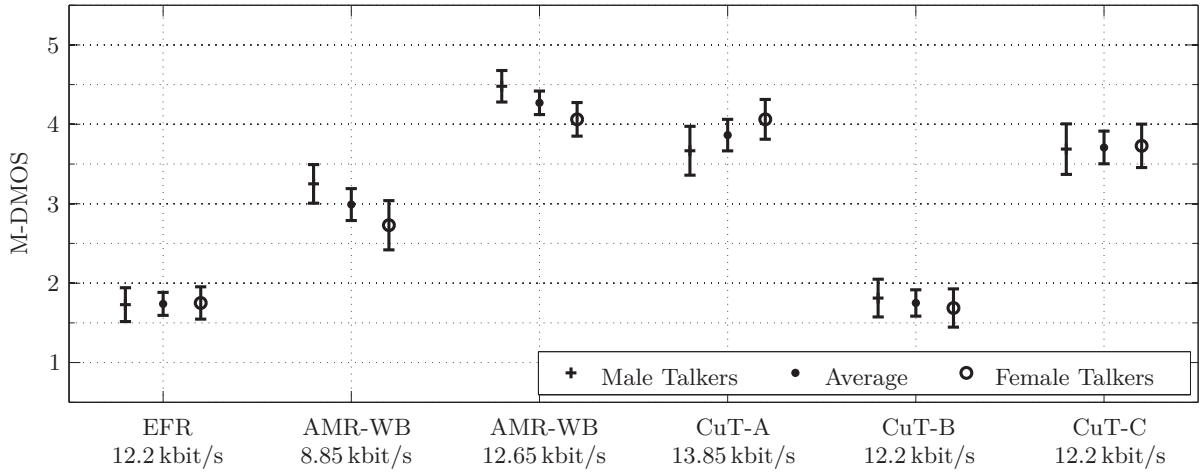


Figure 6.1: Results of the wideband DCR test (95% confidence intervals).

In the tests, the **3GPP EFR** codec serves as the *narrowband anchor condition* because it is widely deployed in current mobile networks, often in the form of the 12.2 kbit/s mode of the 3GPP AMR codec. A relevant quality reference for *wideband* speech is given by the **3GPP AMR-WB** codec [ETSI 2001*b*, Bessette et al. 2002]. The bit rates of 8.85 and 12.65 kbit/s are included in the tests as *reference conditions*.

Test Results

The listening test for the wideband conditions was conducted in two rounds. The tested codecs in the first round were: 3GPP EFR, 3GPP AMR-WB at 12.65 kbit/s, CuT-A, and CuT-C. The remaining conditions (3GPP AMR-WB at 8.85 kbit/s and CuT-B) were assessed in the second round. To maintain the quality anchoring, a few samples for EFR and AMR-WB at 12.65 kbit/s have been included again (although they were not used for the final evaluation). The results are shown in Figure 6.1 with 95% confidence intervals. The M-DMOS scores for male and female talkers are shown separately.

As a *dedicated* wideband codec, 3GPP AMR-WB at 12.65 kbit/s achieved the highest rating among all test conditions while the *embedded* coding approach (CuT-A) is rated only slightly worse on average. For female voices, it is actually equivalent. The fact that a quality gap to AMR-WB at 12.65 kbit/s only exists for male voices indicates that this discrepancy can be attributed to the baseband codec (3GPP EFR) instead of the bandwidth extension part (TDBWE). However, a definite conclusion cannot be drawn from the present experiment.

CuT-B, with its *estimated parameter set*, performs considerably worse than its competitors. In fact, no advantage over narrowband telephony (EFR) could be shown in the present DCR test. Still, a consistent preference of CuT-B over 3GPP EFR was found shown in an informal A-B comparison. These seemingly

conflicting results can be explained by the relatively strict DCR test scenario where the *original* wideband signal is available for a direct comparison. Apparently, the estimated parameters merely convey the *subjective impression* of wideband speech, but the perceivable *differences* in comparison to the reference signal are still obvious. Also, with the high-quality audio equipment provided, residual artifacts (estimation errors) in the synthesized extension band can be better identified than with audio playback through the earpiece of a mobile phone.

The quality of the *steganographic codec* (CuT-C) is almost identical with that of the embedded codec (CuT-A). Obviously, data hiding has very little impact. Only the quality scores for female voices are slightly degraded. This is consistent with the findings of the subjective ABX test in Section 5.5.3.

Compared to narrowband telephony (EFR), which is rated as "slightly annoying," a clearly improved quality could be shown for almost all wideband test conditions. In particular the ratings for AMR-WB at 12.65 kbit/s, CuT-A, and CuT-C are sufficient for "High Definition Telephony." Thereby, in comparison to the traditional, monolithic wideband codec, CuT-A and in particular CuT-C offer much more flexibility and also backwards compatibility.

6.3 Super-Wideband Speech Quality

For the bandwidth extension from wideband (0.05 – 7 kHz) towards super-wideband (0.05 – 14 kHz) *speech* signals, two test setups[1] are considered:

- **CuT-D: Embedded coding** (Chapter 3)
  The super-wideband parameter set for bandwidth extension is determined at the encoder and *quantized* with a bit rate of 4 kbit/s. This information is appended to the 32 kbit/s bitstream of the G.729.1 codec, thus forming an *embedded codec* with two bitstream layers. The sum bit rate is 36 kbit/s. CuT-D is identical with the 36 kbit/s mode of "candidate B" for G.729.1-SWB standardization.

- **CuT-E: Parameter estimation** (Chapter 4)
  Bayesian estimation with Hidden-Markov modeling is used to estimate the super-wideband parameter set based on features of the wideband signal. The estimation algorithm has been configured with a 7-bit codebook and 16 Gaussian mixture components per state. The parameter set and the feature vector are defined in Section 4.4. CuT-E does not increase the bit rate.

In the present listening test, the *reference condition* is the *standardized* super-wideband extension of **ITU-T G.729.1 (Amd. 6)** at a bit rate of 36 kbit/s (which is a direct competitor of CuT-D). A *lower quality anchor* is provided by the 32 kbit/s wideband mode of **ITU-T G.729.1**. In fact, G.729.1 at 32 kbit/s is used to encode the wideband (i.e., baseband) signal in *all four* test conditions.

[1] A steganographic super-wideband codec (corresponding to CuT-C from Section 6.2) has not been included in this test, but preliminary results indicate that this setup is also feasible.

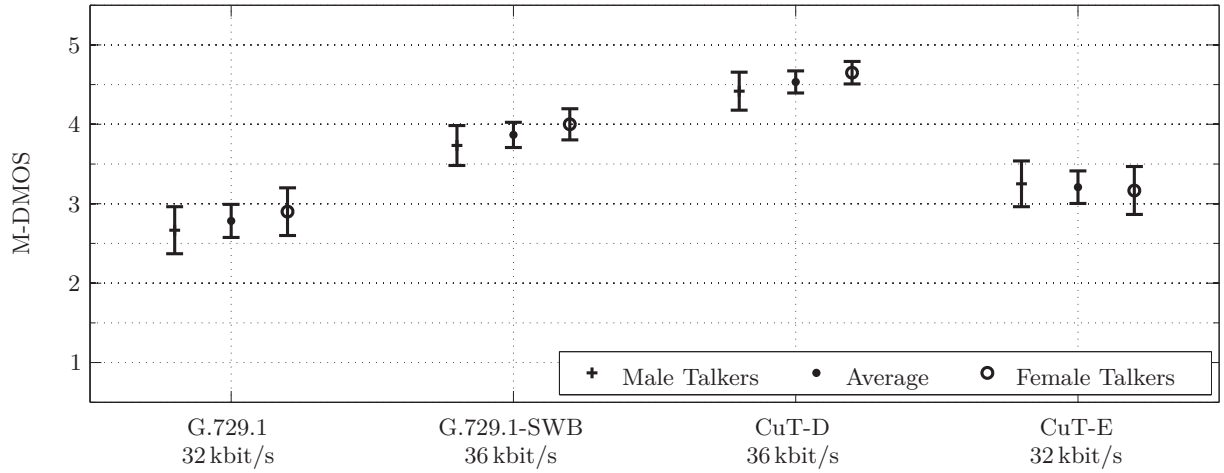


Figure 6.2: Results of the super-wideband DCR test (95% conf. intervals).

Test Results

The results of the listening test for the super-wideband conditions are shown in Figure 6.2 with 95% confidence intervals. Again, the M-DMOS scores for male and female talkers are shown separately.

CuT-D, a.k.a. ITU-T G.729.1-SWB Candidate B, achieves the highest score with a degradation rating between "barely audible" and "inaudible." This finding confirms the excellent super-wideband speech quality of this codec which was already certified by the official ITU-T qualification tests, see Section 3.3.6 and Figure 3.24. Interestingly, the quality score for the *standardized* codec (G.729.1-SWB) is approximately 0.67 M-DMOS *below* that of CuT-D. A more comprehensive comparison of both codecs is provided in Appendix D.

CuT-E, with its *estimated parameter set*, performs considerably worse than the embedded codec variants. However, in contrast to the wideband test results (Figure 6.1), a clear advantage over the wideband condition (G.729.1) could be shown in this test, i.e., artificial bandwidth extension from wideband towards super-wideband frequencies is indeed more reliable than a narrowband to wideband extension, despite its broader extension band.

The "lower anchor" wideband condition (G.729.1) achieves an average degradation rating of "clearly audible, but not annoying." From this, it can be concluded that the 7 – 14 kHz extension band actually adds naturalness and a "sensation of presence" to the speech signal, but, in return, its absence does not lead to a severe quality impairment either. This is in contrast to the wideband listening test (Figure 6.1) where the score for the corresponding *narrowband* anchor condition (EFR) is considerably worse (below 2 M-DMOS).

As a conclusion, a clear improvement over wideband speech (G.729.1) could be shown for all super-wideband test conditions. In particular CuT-D with its average score of 4.53 M-DMOS appears to be well suited for "High Definition Telephony" with a super-wideband bandwidth.

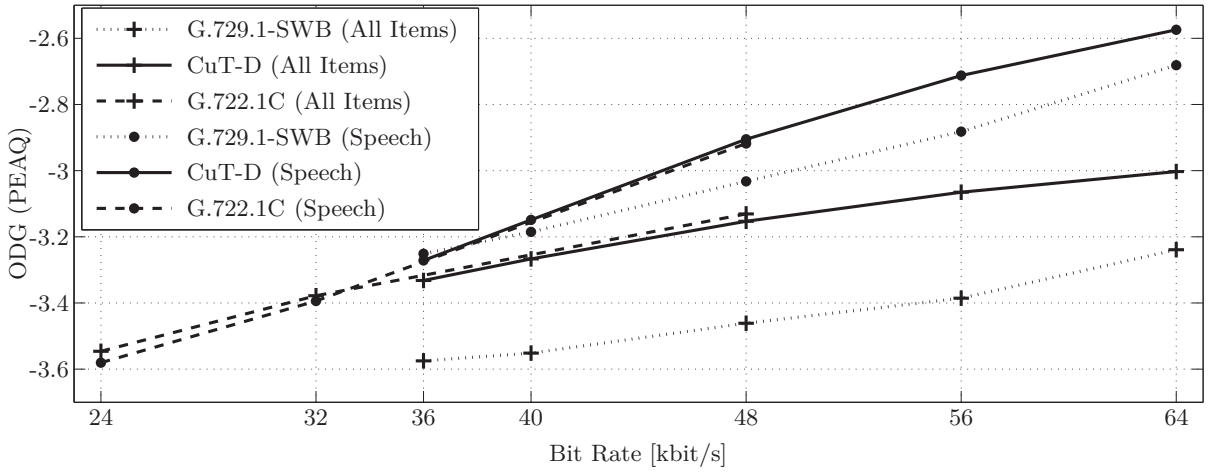


Figure 6.3: Objective audio quality assessment using the PEAQ tool.

6.4 Super-Wideband Audio Quality

The super-wideband bandwidth extension algorithm of Section 3.3 has been explicitly designed for speech *and* audio material. Here, the audio quality is rated using the *objective* PEAQ measure [ITU-R 1998]. The PEAQ scale, given as ODG (Objective Difference Grade), ranges from -4 for the worst quality up to 0 for the best quality. The test conditions are:

- **CuT-D** is tested at bit rates of 36, 40, 48, 56, and 64 kbit/s. Thereby, the low bit rates of 36 and 40 kbit/s can be seen as representative for the proposed bandwidth extension algorithm of Section 3.3. In contrast, the higher bit rates employ classical transform domain vector quantization techniques and the respective PEAQ scores are provided for reference.

- The standardized **ITU-T G.729.1-SWB** codec [Laaksonen et al. 2010] is tested at the same bit rates as CuT-D which facilitates a direct comparison.

- Another super-wideband reference condition is the **ITU-T G.722.1C** codec [ITU-T 1999, Xie et al. 2006] with its bit rates of 24, 32, and 48 kbit/s.

Note that CuT-E is not included in the present test since the artificial bandwidth extension algorithm is *not* suited for general audio and music signals.

Figure 6.3 illustrates the results of the PEAQ evaluation. The measurements have been obtained for the entire EBU SQAM corpus [EBU 1988] (70 items). The scores for the six *speech* items of the database are shown separately.

A clear quality advantage of CuT-D over G.729.1-SWB can be observed for almost all operating conditions. Moreover, CuT-D reaches, unlike G.729.1-SWB, the quality level of G.722.1C. Again, these test results confirm the excellent super-wideband performance of CuT-D which was already found in the ITU-T tests, see Section 3.3.6 and Figure 3.24. A histogram of the PEAQ difference between CuT-D and G.729.1-SWB over all test items is shown in Appendix D.

Chapter 7

Summary

The key feature of a "High Definition" telephone is the reproduction of a wider audio bandwidth which, compared to current "narrowband telephony," brings a much more natural communication experience with a clearly improved speech intelligibility and enhanced listening comfort. Appropriate high quality speech and audio codecs have been developed in the past. However, for a *consequent* introduction of high quality speech and audio transmission in today's heterogeneous communication networks, provisions for *interoperability* and *backwards compatibility* with the installed network infrastructure need to be made.

Therefore, in this thesis, new concepts, methods and algorithms have been studied and devised that facilitate a major audio quality upgrade of *existing* speech communication systems while maintaining backwards compatibility with the installed infrastructure. In a given transmission system, using a well-defined speech codec, interoperability can be provided by techniques for parametric *bandwidth extension*, i.e., the synthesis of missing higher audio frequencies based on a compact parametric description. There are different possibilities to supply the respective parameters to the receiving terminal, as summarized in Table 7.1 at the end of this chapter. Accordingly, the main contributions of this thesis can be categorized into three principal scenarios.

Bandwidth Extension for Embedded Speech and Audio Coding

As a first application scenario, the parameters for bandwidth extension can be *quantized* and appended to the codec bitstream in the form of an "enhancement layer" which can be discarded anywhere in the network without notifying the encoder. Hence, *interoperability* with the original speech codec is maintained, albeit only at the *basic* quality level. In the thesis, two novel bandwidth extension algorithms for embedded speech and audio coding have been described. Both algorithms have been developed in the course of ITU-T standardization projects:

- *Time Domain Bandwidth Extension* (TDBWE) of narrowband speech signals towards the "wideband" frequency range (50 Hz – 7 kHz). This algorithm has been standardized as a part of the recent VoIP codec ITU-T Rec. G.729.1 which extends the widely deployed G.729 narrowband codec. The bit rate for parameter transmission is 1.65 kbit/s.

- *Transform Domain Bandwidth Extension* for wideband *audio* signals towards the "super-wideband" frequency range (50 Hz – 14 kHz). This algorithm has been proposed for standardization in ITU-T where it proved to be the only competitor to fulfill all quality requirements. The bit rate for parameter transmission ranges from 2.7 kbit/s to 4.75 kbit/s.

Several novel signal parametrization methods and synthesis techniques have been included in these algorithms. The main advances can be summarized as follows:

- A new proposal for *temporal envelope control* of synthetic higher audio frequencies has been made. The simple concept of a *temporal gain function* (TGF) based on *subframe gain parameters* not only allows the accurate reproduction of the temporal gain contour of transient as well as stationary signal segments, but it also proved to be extremely versatile in concealing frame erasures that occur in packet-switched networks.

- Several methods for *spectral envelope* modeling and synthesis have been investigated. In the described, practical bandwidth extension algorithms, *subband gains* were computed from a frequency domain representation (DFT or MDCT). These gain parameters facilitate the restoration of the spectral envelope in the time *and* in the transform domain. In this context, the concept of a *filterbank equalizer* (FBE) was applied to the bandwidth extension problem for the first time.

- Elaborate methods to regenerate the *spectral details* of speech *and* audio signals have been devised. For speech, i.e., in the TDBWE algorithm, no explicit parametrization was required. Instead, for generic audio signals, a hybrid algorithm has been proposed which either performs "spectral replication" or "harmonic synthesis," depending on the characteristics of the current signal segment. As a novelty, a concise signal analysis and synthesis could be achieved entirely in the MDCT domain, where previous algorithms usually require an additional, complex-valued frequency transform.

Meanwhile, the transmission of a parametric description of higher audio frequencies within an additional bitstream layer, i.e., *embedded coding*, is a widely applied technique in recent speech and audio coding standards. In fact, a very high quality level can be achieved with this approach which was confirmed by listening tests.

Artificial Bandwidth Extension without Auxiliary Information

If no explicit description of the missing higher audio frequencies is available at all, a certain quality enhancement can, nevertheless, still be obtained with the help of *statistical estimation techniques*. In this scenario, also termed "*Artificial* Bandwidth Extension" (ABWE), the parameters to describe the missing audio frequencies are estimated from the received, band-limited signal alone.

In the thesis, an existing estimation technique, based on Hidden Markov Modeling (HMM), has been applied to the parameter sets of the bandwidth extension algorithms that have been developed for embedded coding, resulting in the following findings:

- Compared to embedded coding, a *reduced* parameter set is usually sufficient for ABWE with statistical estimation. The reason is that the *mutual information* between the received baseband signal and the missing high band parameters is not sufficient to justify the required additional complexity.

- Current statistical estimation methods disregard *estimation errors* that have been produced in *previous* signal frames. As a matter of fact, even a fully *correct* estimation result in the *current* signal frame can produce an unfavorable artifact if the estimated parameter is not consistent with the *previous* (erroneous) estimation result. As a solution, parameter post-processing has been proposed in this thesis.

- The extension of wideband towards super-wideband speech signals (which has not been extensively studied before) yields better and much more consistent estimation results than the typical narrowband to wideband extension scenario. This observation could be substantiated with listening test results.

- As a new application for ABWE techniques, the mitigation of intermittent *bandwidth switchings*, that can occur with embedded codecs, has been identified. During periods of network congestion, in which the *quantized* bandwidth extension parameters could not be received, an *estimated* high band signal is appropriately inserted.

The first applications of artificial bandwidth extension techniques are meanwhile commercially available within certain mobile phones and in automotive applications. Concerning the achievable quality, there is, clearly, a large gap to the embedded coding approach. Nevertheless, a consistent improvement over the bandlimited signal can be confirmed, in particular for the case of a wideband to super-wideband extension. Still, an inherent disadvantage of the ABWE approach is its limitation to *speech* signals.

Bandwidth Extension with Steganographic Parameter Transmission

An attractive compromise between embedded coding techniques and the receiver-based artificial bandwidth extension approach is the *hidden transmission* of the bandwidth extension parameters *within in the standard bitstream* of the baseband codec using *steganographic techniques*. The key advantage of this approach is its *full* backwards compatibility with legacy systems, i.e., the *bitstream format* of the standard speech codec is *not modified*.

In the thesis, three fundamental concepts for data hiding in conjunction with a source encoder have been investigated and evaluated, namely digital watermarking (DWM), bitstream data hiding (BSDH), and joint source coding and data hiding (JSCDH). Thereby, the latter approach was found to be best suited for the present application. With JSCDH, the distortion that is introduced in the speech signal is minimized while the hidden bits can still be *perfectly* reconstructed from the received bitstream.

Consequently, a novel method for JSCDH in state-of-the-art ACELP speech codecs (as used in GSM and UMTS mobile telephony) has been devised. The ACELP data hiding is performed jointly with the analysis-by-synthesis search for the *fixed codebook* of the codec, exploiting the fact that the *standard* ACELP codebook search is *by far* non-exhaustive, i.e., the vast majority of codebook entries is not examined by the standard. Hence, based on a novel *algebraic codebook partitioning* scheme and new *steganographic codebook search* procedure, codebook entries can be taken into account that have been disregarded in the standard (non-steganographic) implementation. The impact of the data hiding operation on the speech quality is therefore very small. For instance, for the GSM EFR codec, very high hidden data rates of up to 2 kbit/s (16% of the codec rate of 12.2 kbit/s) could be achieved without a noteworthy impact on the speech quality. Yet, to further minimize this impact, *variable bit rate data hiding* has been proposed, where, essentially, no measurable quality loss remains.

In a second step, the steganographic version of the GSM EFR codec has been combined with the TDBWE bandwidth extension algorithm, thus forming a fully *backwards compatible wideband codec*. This codec, labeled "Enhanced EFR" (E^2FR) codec, actually achieves a competitive performance, i.e., the achieved wideband speech quality is comparable to that of other wideband codecs. Moreover, link-level simulations for a GSM radio transmission have been conducted, demonstrating the practical relevance of the proposal, at least for *tandem-free* operation in mobile communication networks.

The steganographic transmission of hidden data with the intention of quality enhancement is a quite new proposal. With its relatively high hidden bit rate and the negligible quality impact on legacy systems, the proposed method might be an attractive possibility to upgrade many of today's speech communication systems with enhanced quality (such as wideband or super-wideband telephony) or with other services while maintaining full compatibility with the existing equipment. The method is, in fact, immediately usable in real-world networks by just upgrading the end-user terminals to the new codec version. The ACELP JSCDH technique might even be an attractive option for future standardization efforts. As the JSCDH methods are an integral part of the source encoder, there is the possibility to amend the respective *codec standard* with such functionality. This facilitates a scheduled upgrade of existing communication systems on a large scale.

Conclusion

It can be concluded that, even today, the typical telephony experience is still well characterized with this pointed statement:

"A telephone sounds like a telephone because it is a telephone!"

However, such long-accustomed attitudes of expectation will probably soon belong to the past, mainly because of two contraindications: First, the shift of the current network infrastructure towards the packet-switching paradigm introduces more flexibility into the networks and imposes a less rigid system architecture. New audio codecs are deployed much quicker in such an environment. Second, with new and more advanced methods and algorithms for "High Definition" audio transmission and reproduction, as, i.a., discussed in this thesis, a faster upgrade of—and integration with—the existing, heterogeneous networks becomes feasible.

Table 7.1: Application scenarios of the proposed methods and algorithms in a heterogeneous network.
HD: Device with HD Audio capability. NB: Device without HD Audio capability.
Notation: (sending terminal) → (network) → (receiving terminal)

	HD→HD→HD	HD→NB→HD	NB→HD→HD	NB→NB→HD
dedicated codec (e.g., 3GPP AMR-WB)	HD reproduction is immediately available	transcoding to NB and ABWE in the terminal	ABWE in the network or in the terminal	ABWE in the terminal
embedded codec (e.g., ITU-T G.729.1, see Chapter 3)	HD reproduction is immediately available	drop bitstream extension layer and ABWE in the terminal	ABWE in the network or in the terminal	ABWE in the terminal
steganographic codec, (e.g., E^2FR, see Chapter 5)	HD reproduction is immediately available	HD reproduction is immediately available	ABWE in the network or in the terminal	ABWE in the terminal

Appendix A

Excitation Signal Synthesis in TDBWE

In this appendix, the excitation signal generator of the TDBWE bandwidth extension algorithm, depicted in Figure 3.5, is described in detail. The present time domain excitation generator is designed to regenerate the spectral details in the high frequency band of wideband speech signals (4 – 7 kHz).

Parameters from the CELP Core Layer

The algorithm reuses the following parameters which are transmitted in Layers 1 and 2 of the G.729.1 bitstream, cf. [ITU-T 2006, Massaloux et al. 2007]:

- the integer pitch lag T_0 of the embedded G.729 CELP codec,
- the respective fractional pitch lag $T_{0,\mathrm{frac}}$,
- the energy of the fixed codebook contributions from the core and cascade CELP layers, computed for the current 5 ms subframe according to

$$E_\mathrm{c} = \sum_{k=0}^{39} \left(\hat{g}_\mathrm{c} \cdot c(k) + \hat{g}_\mathrm{enh} \cdot c'(k)\right)^2, \tag{A.1}$$

  where $c(k)$ are the codevector components from the fixed codebook of the core layer CELP codec with its associated quantized gain factor $\hat{g}_\mathrm{c}$, while $c'(k)$ and $\hat{g}_\mathrm{enh}$ are the respective parameters from the cascade CELP layer,

- and the energy of the embedded CELP adaptive codebook contribution for the current 5 ms subframe which is given by

$$E_\mathrm{p} = \sum_{k=0}^{39} \left(\hat{g}_\mathrm{p} \cdot u(k)\right)^2 \tag{A.2}$$

  with the vector components $u(k)$ from the adaptive codebook of the core layer CELP codec and its associated quantized gain factor $\hat{g}_\mathrm{p}$.

Overview

As summarized in Section 3.2.3, the excitation signal generation is structured as follows:

i) estimation of two gains g_v and g_uv for the voiced and unvoiced contributions to the excitation signal $\hat{u}_\mathrm{hb}(k)$,

ii) pitch lag post-processing,

iii) production of the voiced contribution,

iv) production of the unvoiced contribution, and

v) lowpass filtering.

These individual steps are described in the following.

Description of the Algorithm

i) Estimation of gains for the voiced and unvoiced contributions. First, to get an initial estimate of the "harmonics-to-noise" ratio, an instantaneous energy ratio ξ of the adaptive codebook and fixed codebook (including the cascade CELP fixed codebook) contributions is computed for each subframe:

$$\xi = \frac{E_\mathrm{p}}{E_\mathrm{c}}. \tag{A.3}$$

In order to reduce the adaptive-to-fixed codebook power ratio in case of unvoiced sounds, a "Wiener Filter" characteristic is applied to ξ:

$$\xi_\mathrm{post} = \xi \cdot \frac{\xi}{1+\xi}. \tag{A.4}$$

This leads to more consistent unvoiced sounds. Finally, the gains for the voiced and unvoiced contributions to $\hat{u}_\mathrm{hb}(k)$ can be determined. Therefore, an intermediate voiced gain g'_v is calculated:

$$g'_\mathrm{v} = \sqrt{\frac{\xi_\mathrm{post}}{1+\xi_\mathrm{post}}}. \tag{A.5}$$

With a 2-tap gliding average filter, g'_v is smoothed to obtain the final voiced gain

$$g_\mathrm{v} = \sqrt{\frac{1}{2}\left({g'_\mathrm{v}}^2 + {g'_\mathrm{v,old}}^2\right)}, \tag{A.6}$$

where $g'_\mathrm{v,old}$ is the intermediate voiced gain according to (A.5) from the preceding subframe. The averaging of the squared values favors a fast increase of g_v in case of an unvoiced to voiced transition. To satisfy the constraint $g_\mathrm{v}^2 + g_\mathrm{uv}^2 = 1$, the unvoiced gain is now given by

$$g_\mathrm{uv} = \sqrt{1 - g_\mathrm{v}^2}. \tag{A.7}$$

ii) Pitch lag post-processing. The production of a consistent pitch structure within the excitation signal $\hat{u}_{\mathrm{hb}}(k)$ requires a good estimate of the fundamental speech frequency F_0 of the speech production process or of its inverse, the pitch lag t_0. Within Layer 1 of the G.729.1 bitstream, the integer and fractional pitch lag values T_0 and $T_{0,\mathrm{frac}}$ (cf. [ITU-T 1996*b*]) are available for the four 5 ms subframes of the current frame. The present estimation method for t_0 is based on these parameters. It is worth noting that the aim of the encoder-side pitch search procedure in the CELP layer is to find the pitch lag values T_0 and $T_{0,\mathrm{frac}}$ which minimize the power of the Long Term Prediction (LTP) residual signal. Consequently, the LTP pitch lag is not necessarily identical with t_0, which is a requirement for a concise synthetic reproduction of voiced speech components. The most typical deviations are pitch-doubling and pitch-halving errors, i.e., the frequency corresponding to the LTP lag is half or double that of the original fundamental speech frequency. In particular, pitch-doubling (-tripling, etc.) errors need to be avoided. Hence, the following post-processing of the LTP lag information (T_0 and $T_{0,\mathrm{frac}}$) is used.

First, the LTP pitch lag for an oversampled time-scale is reconstructed from T_0 and $T_{0,\mathrm{frac}}$. Because the fractional resolution of the pitch lag in the G.729.1 CELP layer is as precise as 1/3 of a sample, the oversampled lag amounts to $3\,T_0 + T_{0,\mathrm{frac}}$. Then, an additional factor of 2 is considered such that an enhanced resolution (see (A.12)) can be represented:

$$t_{\mathrm{LTP}} = 2 \cdot (3\,T_0 + T_{0,\mathrm{frac}}). \tag{A.8}$$

The (integer) factor between the currently observed LTP lag t_{LTP} and the post-processed pitch lag of the preceding subframe $t_{\mathrm{post,old}}$ (see (A.11)) is calculated by[1]

$$\rho = \left\lfloor \frac{t_{\mathrm{LTP}}}{t_{\mathrm{post,old}}} + 0.5 \right\rfloor . \tag{A.9}$$

If the factor ρ falls into the range $2, \ldots, 4$, a relative error is evaluated:

$$e = 1 - \frac{t_{\mathrm{LTP}}}{\rho \cdot t_{\mathrm{post,old}}}. \tag{A.10}$$

If the magnitude of this relative error is below a threshold of $\epsilon = 0.1$, it is assumed that the current LTP lag is the result of a beginning pitch-doubling (-tripling, -quadruplication) error phase. Thus, the pitch lag is corrected with a division by the integer factor ρ, thereby producing a continuous pitch lag behavior w.r.t. the previous pitch lags:

$$t_{\mathrm{post}} = \begin{cases} \left\lfloor \frac{t_{\mathrm{LTP}}}{\rho} + 0.5 \right\rfloor & \text{if } |e| < \epsilon, \rho > 1, \rho < 5 \\ t_{\mathrm{LTP}} & \text{otherwise.} \end{cases} \tag{A.11}$$

[1] $\lfloor x \rfloor$ denotes the highest integer number not greater than x.

A gliding average filter with 2 taps is applied to t_{post}:

$$t_{\mathrm{p}} = \frac{1}{2}\left(t_{\mathrm{post,old}} + t_{\mathrm{post}}\right). \tag{A.12}$$

Note that this gliding average leads to a virtual precision enhancement from a resolution of 1/3 to 1/6 of a sample. Finally, the post-processed pitch lag t_{p} is decomposed into its integer and fractional parts

$$t_{0,\mathrm{int}} = \left\lfloor \frac{t_{\mathrm{p}}}{6} \right\rfloor \quad \text{and} \quad t_{0,\mathrm{frac}} = t_{\mathrm{p}} - 6 \cdot t_{0,\mathrm{int}}. \tag{A.13}$$

iii) Production of the voiced contribution. The voiced components $\hat{u}^{\mathrm{v}}_{\mathrm{hb}}(k)$ of the excitation signal are represented as shaped and weighted glottal pulses. In the following, these pulses are indexed by the global "counter" p. Hence, $\hat{u}^{\mathrm{v}}_{\mathrm{hb}}(k)$ is produced by overlap-add of single pulse contributions for the current 5 ms subframe:

$$\hat{u}^{\mathrm{v}}_{\mathrm{hb}}(k) = \sum_{p:\ 0 \le k - k^{[p]}_{\mathrm{p,int}} \le 56} g^{[p]}_{\mathrm{p}} \cdot P_{k^{[p]}_{\mathrm{p,frac}}}\left(k - k^{[p]}_{\mathrm{p,int}}\right), \tag{A.14}$$

where $g^{[p]}_{\mathrm{p}}$ is the *gain* factor for each pulse, $k^{[p]}_{\mathrm{p,int}}$ is the pulse *position*, and $P_i(k)$ is the i-th *pulse shape prototype*. The selection of the prototype depends on a "fractional pulse position" $i = k^{[p]}_{\mathrm{p,frac}}$. These parameters are derived in the following.

The post-processed pitch lag parameters $t_{0,\mathrm{int}}$ and $t_{0,\mathrm{frac}}$ determine the pulse spacing, hence, the pulse positions according to

$$k^{[p]}_{\mathrm{p,int}} = k^{[p-1]}_{\mathrm{p,int}} + t_{0,\mathrm{int}} + \left\lfloor \frac{k^{[p-1]}_{\mathrm{p,frac}} + t_{0,\mathrm{frac}}}{6} \right\rfloor, \tag{A.15}$$

where $k^{[p]}_{\mathrm{p,int}}$ is the (integer) position of the current pulse and $k^{[p-1]}_{\mathrm{p,int}}$ is the (integer) position of the previous pulse. The fractional part of the pulse position

$$k^{[p]}_{\mathrm{p,frac}} = k^{[p-1]}_{\mathrm{p,frac}} + t_{0,\mathrm{frac}} - 6 \cdot \left\lfloor \frac{k^{[p-1]}_{\mathrm{p,frac}} + t_{0,\mathrm{frac}}}{6} \right\rfloor \tag{A.16}$$

serves as an index for the pulse shape selection. The prototype pulse shapes with $i \in \{0, \ldots, 5\}$ and $k \in \{0, \ldots, 56\}$ are taken from a lookup table which is plotted in Figure A.1.

The pulse shape prototypes $P_i(k)$ in this lookup table are filtered and resampled versions of a wideband (16 kHz) pulse from a "typical" voiced speech segment. The segment was selected for its specific spectral characteristics which avoid an "overvoicing" of the excitation (cf. discussion below). Since a sampling frequency of 8 kHz and a resolution of 1/6 of a sample is targeted for the given application, the selected pulse has been upsampled to 48 kHz first. The six final pulse shapes $P_i(k)$ have then been obtained by applying the following operations:

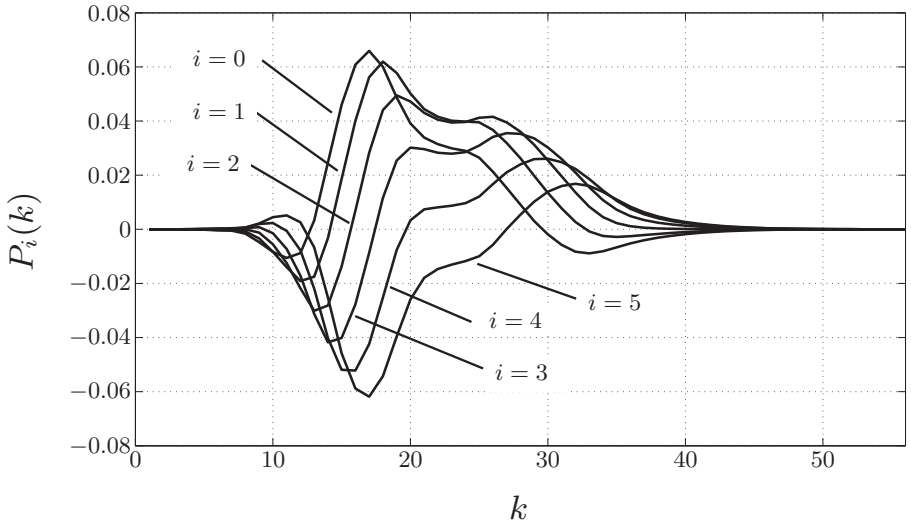


Figure A.1: Pulse shape lookup table for the *voiced* contribution $\hat{u}^{\mathrm{v}}_{\mathrm{hb}}(k)$ to the synthetic excitation signal $\hat{u}_{\mathrm{hb}}(k)$.

- lowpass filtering and decimation by a factor of 3 (with 3 subsampling offsets),
- highpass filtering and decimation by a factor of 2 (with 2 subsamp. offsets),
- and spectral mirroring, i.e., multiplication by $(-1)^k$.

Note that spectral mirroring $(-1)^k$ may give two different results depending on the starting position of the pulse (even or odd sample index). This fact is accounted for in the pulse gain calculation (cf. first factor in (A.17)). The gain factors $g_{\mathrm{p}}^{[p]}$ for the individual pulses (with index p) are, apart from the position dependent sign inversion, derived from the voiced gain parameter g_{v} and from the pitch lag parameters:

$$g_{\mathrm{p}}^{[p]} = \left(2 \cdot \mathrm{even}\left(k_{\mathrm{p,int}}^{[p]}\right) - 1\right) \cdot g_{\mathrm{v}} \cdot \sqrt{6\, t_{0,\mathrm{int}} + t_{0,\mathrm{frac}}}. \tag{A.17}$$

Here, the square root ensures that the varying pulse spacing does not have an impact on the resulting signal energy. The function even$(\cdot)$ returns 1 if the argument is an even integer number and 0 otherwise.

With the design described above, the full sub-sample resolution of the pitch lag information can be utilized by a simple pulse shape selection. Further, the pulse shapes exhibit a certain spectral shaping which ensures smoothly attenuated higher frequency components of the voiced excitation. This avoids a high frequency "overvoicing." Additionally, compared to unit pulses, the applied pulse shapes result in a strongly reduced crest factor of the excitation signal which leads to an improved subjective quality.

iv) Production of the unvoiced contribution. The unvoiced contribution $\hat{u}_{\mathrm{hb}}^{\mathrm{uv}}(k)$ for each 5 ms subframe is produced using with a white noise generator:

$$\hat{u}_{\mathrm{hb}}^{\mathrm{uv}}(k) = g_{\mathrm{uv}} \cdot n(k), \tag{A.18}$$

where $k \in \{0, \ldots, 39\}$. The random generator is identical with the generator used in the G.729 codec. It produces a signal of unit variance.

v) Lowpass filtering. With the voiced and unvoiced contributions $\hat{u}_{\mathrm{hb}}^{\mathrm{v}}(k)$ and $\hat{u}_{\mathrm{hb}}^{\mathrm{uv}}(k)$, the final excitation signal $\hat{u}_{\mathrm{hb}}(k)$ is obtained by lowpass filtering of $\hat{u}_{\mathrm{hb}}^{\mathrm{v}}(k) + \hat{u}_{\mathrm{hb}}^{\mathrm{uv}}(k)$. The 3 kHz lowpass filter is identical with the pre-processing lowpass filter for the high band signal as shown in Figure 3.2.

Appendix B

Data Hiding and Source Coding (DWM, BSDH, JSCDH)

In this appendix, as a supplement to Section 5.2, the three concepts for combined data hiding and source coding, i.e., Digital Watermarking (DWM), BitStream Data Hiding (BSDH), and Joint Source Coding and Data Hiding (JSCDH), shall be analyzed and compared. The respective block diagrams for the three systems are illustrated in Figure 5.3.

In particular, the distortion penalties of DWM, BSDH, and JSCDH over mere source coding (as shown in Figure 5.5) are derived analytically. In addition, the respective error probabilities for the hidden bits that already occur *without any channel noise* are given.

Note that the focus here is to highlight the fundamental differences between the three approaches in a comprehensible fashion. Therefore, the present analysis is limited to *scalar* quantization methods.

Preliminaries

The (infinite) codebook $\mathcal{C} = \Delta\mathbb{Z}$ of a uniform scalar quantizer with stepsize Δ shall be used for the *source coding* (SC) components of all three systems under consideration. The resulting quantization distortion is $D_{\mathrm{SC}} = \Delta^2/12$.

Instead, for the *data hiding* components, $M > 1$ disjoint sub-codebooks $\mathcal{C}_m = \Delta\,(M\mathbb{Z}+m)$ are used, where $m \in \{0, \ldots, M-1\}$ is the message to be hidden in the scalar output value $\tilde{x}$. The codebooks $\mathcal{C}_m$ can be immediately applied in the JSCDH and BSDH systems. However, a fair comparison of all *three* systems requires modified DWM codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ that do not exploit intricate knowledge of the source codebook $\mathcal{C}$. This independence is achieved with a *dithered quantization* approach for the DWM system. The respective details are described in the DWM section of this appendix.

For all systems, the input signal x is assumed to be uniformly distributed within each quantization interval and to have a sufficiently large variance $\sigma_x^2 \gg \sigma_{x-\tilde{x}}^2$. The coded and transmitted signal (containing the hidden message m) is $\tilde{x} \in \mathcal{C}$. The average total distortion per sample that is then incurred by the different systems is computed as $D_{(\cdot)} \doteq \mathrm{E}\left\{(x - \tilde{x}_{(\cdot)})^2\right\}$.

Joint Source Coding and Data Hiding (JSCDH)

With $\mathcal{C}_m \subset \mathcal{C}$, the scalar JSCDH system is equivalent to a quantizer with an increased stepsize of $M\Delta$. Therefore, the incurred total distortion is:

$$D_{\mathrm{JSCDH}} = \int_{-\frac{M\Delta}{2}}^{\frac{M\Delta}{2}} \frac{u^2}{M\Delta}\,\mathrm{d}u = \frac{\Delta^2}{12}M^2, \tag{B.1}$$

i.e., the distortion penalty compared to the source coding system can be quantified as $D_{\mathrm{JSCDH}}/D_{\mathrm{SC}} = M^2 = 2^{2R_{\mathrm{DH}}}$. This ratio is shown in Figure 5.5(a) in decibel.

As the JSCDH codewords are, by construction, elements of the source codebook $\mathcal{C}$, and, as no further channel noise is assumed, the probability of a decoding error is exactly zero, i.e., $P_{\mathrm{e,JSCDH}} = 0$.

Bitstream Data Hiding (BSDH)

For the BSDH system, a natural binary bit representation of the quantization levels is considered, i.e., if the quantization level $i \cdot \Delta$ is represented by the number $b_i = i$ in binary form, then $(i+1) \cdot \Delta$ is represented by $b_{i+1} = i + 1$.

Now there are two possible realizations of bitstream data hiding. The first is a simple *replacement* of the $\mathrm{ld}\,M$ least significant bits (LSBs) of the bit pattern that is associated with the source coded value $\hat{x}$ by the message m (also in binary form). The PDF of the resulting overall noise is then given as the convolution of the source coding noise PDF $p_{\mathrm{n,SC}}(u)$ with the PDF of the *additional* BSDH noise $p_{\mathrm{n,BSDH}}(u)$. The BSDH noise is discrete because only integer multiples of Δ may be added to or subtracted from the source coded signal $\hat{x}$ depending on the particular message m. Therefore, the convolution leads to an overall noise PDF that is uniform in the (asymmetric) interval $[-\Delta(i+\frac{1}{2}), \Delta(M-\frac{1}{2}-i)]$. The interval boundaries depend on the actual LSBs $i \in \{0, \ldots, M-1\}$ of the source coded value $\hat{x}$. Hence, the mean total distortion can be obtained by averaging over i:

$$D_{\mathrm{BSDH,LSB,avg}} = \frac{1}{M}\sum_{i=0}^{M-1} \int_{-(i+\frac{1}{2})\Delta}^{(M-\frac{1}{2}-i)\Delta} \frac{u^2}{M\Delta}\,\mathrm{d}u = \frac{\Delta^2}{12}(2M^2-1). \tag{B.2}$$

When compared with the JSCDH distortion (B.1), an additional distortion penalty of $D_{\mathrm{BSDH,LSB,avg}}/D_{\mathrm{JSCDH}} = 2 - 1/M^2 = 2 - 2^{-2R_{\mathrm{DH}}}$ emerges. This is displayed in Figure 5.5(b) using the "×" marker.[1]

In addition to the average BSDH distortion of (B.2), also the worst case limit shall be considered. The maximum distortion level is incurred for a maximally

[1] The individual markers are placed at *even* values of M.

asymmetric BSDH noise PDF, i.e., for $i = 0$:

$$D_{\mathrm{BSDH,LSB,max}} = \int_{-\frac{\Delta}{2}}^{(M-\frac{1}{2})\Delta} \frac{u^2}{M\Delta}\,\mathrm{d}u = \frac{\Delta^2}{12}(4M^2 - 6M + 3). \tag{B.3}$$

The corresponding additional distortion penalty over the JSCDH system, shown with the "$\circ$" marker, is $D_{\mathrm{BSDH,LSB,max}}/D_{\mathrm{JSCDH}} = 4 - 6/M + 3/M^2 = 4 - 6 \cdot 2^{-R_{\mathrm{DH}}} + 3 \cdot 2^{-2R_{\mathrm{DH}}}$.

The second variant of the BSDH system is actually the *best* case that can be achieved over all possible values of the LSBs i of $\hat{x}$. This can be interpreted as a *requantization* (RQ) of the source coded value $\hat{x}$ with the appropriate data hiding sub-codebook $\mathcal{C}_m$. The effective noise PDF is symmetric for odd values of M, but slightly asymmetric for even values of M. Hence, the total distortion is:

$$D_{\mathrm{BSDH,RQ}} = \begin{cases} \int_{-\frac{M\Delta}{2}}^{\frac{M\Delta}{2}} \frac{u^2}{M\Delta}\,\mathrm{d}u = \frac{\Delta^2}{12}M^2 & \text{for} \quad M \quad \text{odd} \\ \int_{\frac{1-M}{2}\Delta}^{\frac{1+M}{2}\Delta} \frac{u^2}{M\Delta}\,\mathrm{d}u = \frac{\Delta^2}{12}(M^2+3) & \text{for} \quad M \quad \text{even.} \end{cases} \tag{B.4}$$

For the latter case, which is more relevant in practice, Figure 5.5(b) shows the corresponding additional distortion penalty $D_{\mathrm{BSDH,RQ}}/D_{\mathrm{JSCDH}} = 1 + 3/M^2 = 1 + 3 \cdot 2^{-2R_{\mathrm{DH}}}$ with a "+" marker. Note that $D_{\mathrm{BSDH,RQ}}$ is actually close (or for odd M even identical) to the JSCDH distortion (B.1). However, intricate knowledge of the utilized source codebook $\mathcal{C}$ and of the applied bit mapping is required to apply requantization. Still, JSCDH is not only slightly better in terms of overall distortion, but it is also less complex and therefore preferable (if feasible in the considered system).

Since BSDH directly manipulates the bitstream, the probability of a decoding error for both variants (LSB and RQ) is, as in the JSCDH case, $P_{\mathrm{e,BSDH}} = 0$.

Digital Watermarking (DWM)

The analysis of the DWM system is not as straight forward as for other approaches. For DWM, to facilitate a *fair* comparison of all *three* systems, it is necessary to use the same *source* codebook $\mathcal{C}$ as for scalar JSCDH and BSDH. However, the scalar data hiding codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ may be chosen arbitrarily. Therefore, as DWM is required to be unaware of the detailed structure of $\mathcal{C}$, codebook randomization in the form of *dithered quantization* is applied in the following analysis. Dithering is in fact a standard technique in quantization based watermarking, cf. [Chen & Wornell 2001, Eggers et al. 2003]. The quantization rule for a dithered scalar quantizer with stepsize Δ is $\hat{x} = \mathcal{Q}_\Delta(x - d) + d$. The dither sequence $\{d\}$ must be

uniformly distributed over the interval $[-\Delta/2, \Delta/2]$ and independent of x. This kind of dithering enforces a uniform quantization error PDF within the same range of values, regardless of the input distribution.

The stepsize for the DWM codebooks $\mathcal{C}_m^{\mathrm{DWM}}$ is related to the source code stepsize as $\Delta_{\mathrm{DWM}} = M\rho\Delta$ with a real-valued shrinkage/expansion factor ρ. The overall noise PDF is again asymmetric with a variable shift of v. Averaging the total distortion over all possible shift values gives:

$$D_{\mathrm{DWM,avg}} = \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta} \int_{-\frac{M\rho\Delta}{2}+v}^{\frac{M\rho\Delta}{2}+v} \frac{u^2}{M\rho\Delta}\,\mathrm{d}u\,\mathrm{d}v = \frac{\Delta^2}{12}(\rho^2M^2+1). \tag{B.5}$$

The maximum distortion value is obtained for a PDF shift of $v = \Delta/2$:

$$D_{\mathrm{DWM,max}} = \int_{\frac{1-M\rho}{2}\Delta}^{\frac{M\rho+1}{2}\Delta} \frac{u^2}{M\rho\Delta}\,\mathrm{d}u\,\mathrm{d}v = \frac{\Delta^2}{12}(\rho^2M^2+3) \tag{B.6}$$

which, for $\rho = 1$, is identical to the result for BSDH with requantization (B.4). In the best case, i.e., for $v = 0$, the distortion is:

$$D_{\mathrm{DWM,min}} = \int_{\frac{-M\rho\Delta}{2}}^{\frac{M\rho\Delta}{2}} \frac{u^2}{M\rho\Delta}\,\mathrm{d}u\,\mathrm{d}v = \frac{\Delta^2}{12}\rho^2M^2 \tag{B.7}$$

which, for $\rho = 1$, is identical to the JSCDH performance.

With the DWM system, decoding errors may occur if the source coding noise $|\tilde{x} - \tilde{x}_{\mathrm{DWM}}|$ exceeds $\rho\Delta$ which designates the stepsize of the *union* of all DWM codebooks $\mathcal{C}_m^{\mathrm{DWM}}$. As $\max|\tilde{x} - \tilde{x}_{\mathrm{DWM}}| = \Delta$, $P_{\mathrm{e,DWM}} = 0$ if $\rho > 1$. If $\rho \leq 1$, the average error probability is:

$$P_{\mathrm{e,DWM}} = 2 \cdot \int_{\frac{\rho\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta}\,\mathrm{d}u = 1-\rho \quad \text{for} \quad \frac{1}{2M-1} \leq \rho \leq 1. \tag{B.8}$$

The lower bound on ρ in (B.8) is due to the fact that a very large source coding noise (larger than $(M-\frac{1}{2})\rho\Delta$) leads to a correctly decoded message m again. Yet, this case is not relevant in practice, as, e.g., for $M = 2$ and $\rho = \frac{1}{2M-1}$, the decoding error probability is already as high as $P_{\mathrm{e,DWM}} = \frac{2}{3}$.

There are in fact two interesting settings for the parameter ρ. First, $\rho = 1$ guarantees an error-free message decoding, but a small distortion penalty compared

to JSCDH has to be taken as analyzed in (B.5) – (B.7). Second, to avoid an additional distortion penalty, $D_{\mathrm{DWM,avg}} \stackrel{!}{=} D_{\mathrm{JSCDH}}$ can be enforced. This, in turn, leads to a non-zero error probability of $P_{\mathrm{e,DWM}} = 1-\rho = 1-\sqrt{1-\frac{1}{M^2}}$ (if $M > 1$).

The "Δ" markers in Figure 5.5(b) show the additional average distortion penalty compared to the JSCDH system for the (error-free) case of $\rho = 1$, i.e., $D_{\mathrm{DWM,avg}}/D_{\mathrm{JSCDH}} = 1 + 1/M^2 = 1 + 2^{-2R_{\mathrm{DH}}}$. Note that the stepsize Δ of the source codebook $\mathcal{C}$ must be known to the watermarking unit in this case. Any mismatch will either unnecessarily increase the distortion penalty or cause decoding errors.

Appendix C

Data Hiding Modes for 3GPP EFR

As a supplement to Section 5.5.2, this appendix summarizes the variants of steganographic codebook search algorithm for the 3GPP EFR codec that allow to hide data rates from 2 kbit/s down to 0.2 kbit/s in the bitstream of the codec.

Codebook Partitioning

In addition to the full (unrestricted) track set $\mathcal{T}_t$ with its eight pulse position candidates, two different restricted sets with either *two* or *four* pulse position candidates are used in the variants of the steganographic codebook search algorithm. They are defined as follows:

- For a *two*-bit message $m_t \in \{0, \ldots, 3\}$ to be embedded in a track $\mathcal{T}_t$ with index t, *two* out of eight pulse position candidates remain, i.e., the restricted track set $\mathcal{T}_t^{m_t}$ is defined as:

  $$\mathcal{T}_t^{m_t} = \Big\{5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus m_t\big) + t, 5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus (m_t + 4)\big) + t\Big\}.$$

  This partitioning is, for instance, used for data hiding with 2 kbit/s. It is thus identical to the description from Section 5.5.2, see e.g., Table 5.4.

- For a *one*-bit message $m_t \in \{0, 1\}$ to be embedded in a track $\mathcal{T}_t$ with index t, *four* out of eight pulse position candidates remain, i.e., the restricted track set $\mathcal{T}_t^{m_t}$ is defined as:

  $$\begin{aligned}\mathcal{T}_t^{m_t} = \{\, &5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus m_t\big) + t,\ 5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus (m_t + 2)\big) + t, \\ &5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus (m_t + 4)\big) + t,\ 5 \cdot \mathcal{G}^{-1}\big(\mathcal{G}(i_t) \oplus (m_t + 6)\big) + t\,\}.\end{aligned}$$

In both variants, i_t designates the index of the (unrestricted) *first* pulse in track $\mathcal{T}_t$. To obtain the various bit rates for steganography, the described track restrictions are combined in different ways. This is summarized in the table below. Thereby, the notation $|\mathcal{T}_t^{m_t}|$ designates the cardinality of the set $\mathcal{T}_t^{m_t}$ which is equivalent to the number of valid pulse position candidates.

For example for $R_{\mathrm{DH}} = 1.65$ kbit/s, three two-bit messages m_0, m_1, and m_2 are embedded in tracks $\mathcal{T}_0$, $\mathcal{T}_1$, and $\mathcal{T}_2$, i.e., only two valid pulse positions are considered within the restricted pulse position sets $\mathcal{T}_0^{m_0}$, $\mathcal{T}_1^{m_1}$, and $\mathcal{T}_2^{m_2}$. For track $\mathcal{T}_4$, four pulse position candidates are examined instead, i.e., a one-bit message m_4 is embedded therein. The data hiding scheme for track $\mathcal{T}_3$ is actually switched

depending on the specific 5 ms subframe which is currently being encoded.[1] A one-bit message $m_3 \in \{0, 1\}$ is embedded in subframes 1 – 3 while a two-bit message $m_3 \in \{0, \ldots, 3\}$ is embedded in subframe 4. The total hidden bit rate can therefore be computed as follows: $(3 \cdot 2 + \frac{3}{4} \cdot 1 + \frac{1}{4} \cdot 2 + 1)\,\text{bit}/5\,\text{ms} = 1.65\,\text{kbit/s}$.

R_{DH}	$\lvert\mathcal{T}_0^{m_0}\rvert$	$\lvert\mathcal{T}_1^{m_1}\rvert$	$\lvert\mathcal{T}_2^{m_2}\rvert$	$\lvert\mathcal{T}_3^{m_3}\rvert$	$\lvert\mathcal{T}_4^{m_4}\rvert$
2 kbit/s	2	2	2	2	2
1.8 kbit/s	2	2	2	2	4
1.65 kbit/s	2	2	2	4 (subframe 1–3) 2 (subframe 4)	4
1.2 kbit/s	2	4	4	4	4
0.8 kbit/s	4	4	4	4	8
0.4 kbit/s	4	4	8	8	8
0.2 kbit/s	4	8	8	8	8

Steganographic Codebook Search

For convenience, the codebook search schedule of Section 5.5.2 (see page 141), which has originally been designed for the maximum hidden bit rate of $R_{\text{DH}} = 2\,\text{kbit/s}$, is reused for the lower hidden bit rates.

Note that, as a consequence of these identical search schedules, the search complexity gradually increases with lower bit rates. This complexity increase can be avoided with dedicated search schedules that are particularly tailored to the specific hidden bit rate.

[1] The 3GPP EFR codec divides each 20 ms speech frame into four 5 ms subframes.

Appendix D

Additional Test Results

For reference and for comparison with the results of the subjective listening tests from Chapter 6, *objective* quality scores for CuT-A – CuT-E have been measured. For the wideband speech quality, PESQ measurements have been conducted. For the super-wideband case, the PEAQ tool has been used instead.

In addition, CuT-D (i.e., "Candidate Codec B" for the G.729.1-SWB standardization [Geiser et al. 2009]) is directly contrasted with the standardized version of the codec [Laaksonen et al. 2010].

Wideband Speech Quality (PESQ)

The wideband version of the ITU-T PESQ tool [ITU-T 2005] has been used to measure the wideband speech quality of CuT-A – CuT-C and of the respective reference conditions (see Section 6.2). The results are shown in Figure D.1.

In contrast to Figure 6.1, also the ITU-T G.729.1 codec is included here as an additional wideband reference. The average WB-PESQ scores and the standard deviations have been computed for the entire NTT database [NTT 1994] at an input level of −26 dBov [ITU-T 1993*b*].

In general, the PESQ scores are concentrated between values of 3 and 4 on the MOS-LQO scale. The general trend for CuT-A and CuT-C is in line with the outcome of the subjective test, only the rating for AMR-WB at 12.65 kbit/s is somewhat worse (equivalent to CuT-C). Its large standard deviation points at the quality discrepancy between male and female voices that has been already found in the subjective test. The G.729.1 codec at 14 kbit/s, which is based on the same bandwidth extension technique as CuT-A, indeed achieves the same performance as CuT-A.

Furthermore, it can also be observed tat the WB-PESQ scores for the narrowband EFR codec and for CuT-B surprisingly high. However, the usefulness of the wideband PESQ measure to assess narrowband speech coding (EFR) as well as artificial bandwidth extension techniques (CuT-B) is debatable.

Super-Wideband Speech Quality (PEAQ)

The super-wideband speech quality of CuT-D, CuT-E, and of the respective reference conditions has been assessed with the PEAQ tool [ITU-R 1998]. The results in Figure D.2 show the mean quality scores and the standard deviation which have been computed for a super-wideband version of the NTT database [NTT 1994] at an input level of −26 dBov [ITU-T 1993*b*].

The principal observation in Section 6.3, i.e., the quality advantage of CuT-D over G.729.1-SWB, is also evident in this test. Similar to the wideband test from Figure D.1, the lower anchor condition (G.729.1 at 32 kbit/s) receives a comparatively high quality rating (better than CuT-E).

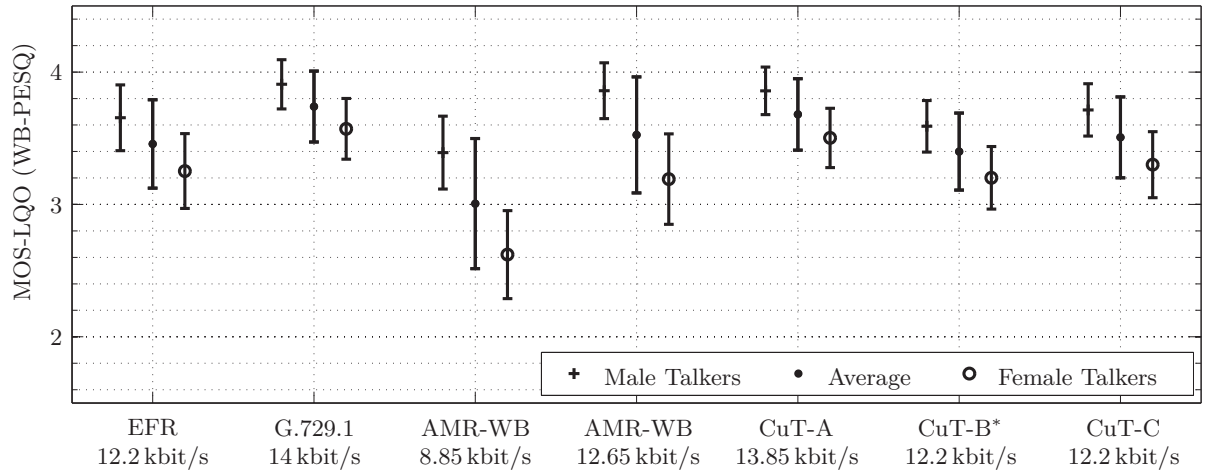


Figure D.1: Quality assessment with WB-PESQ (mean and std. dev.)
*Training material has been excluded from the test corpus for CuT-B.

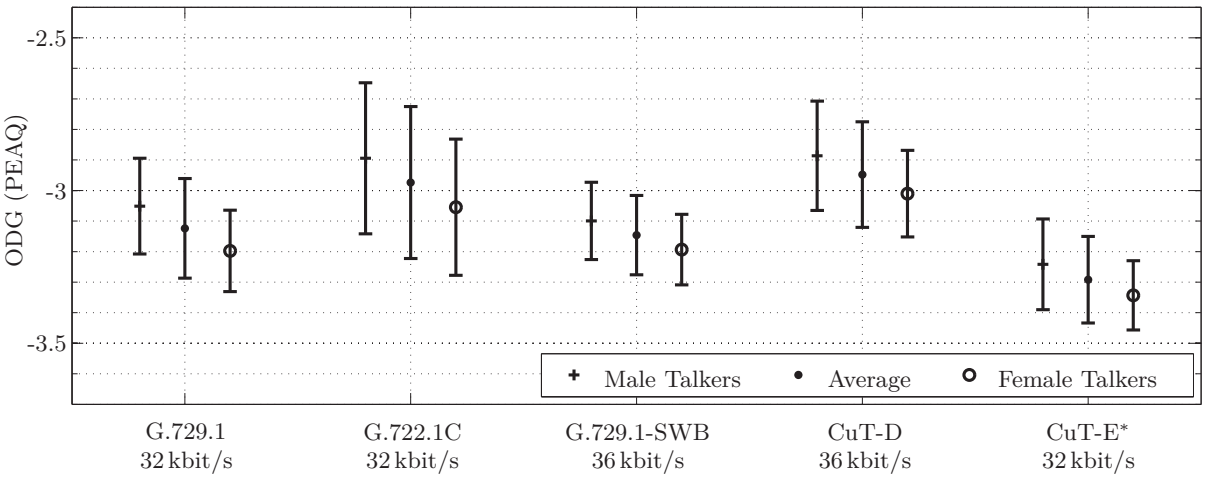


Figure D.2: Quality assessment with PEAQ (mean and std. dev.)
*Training material has been excluded from the test corpus for CuT-E.

Comparison of CuT-D and ITU-T G.729.1-SWB

Supplementing the test results of Section 6.4, Figure D.3 shows a histogram of the PEAQ score difference between CuT-D and ITU-T Rec. G.729.1 Amd. 6. The difference score is defined as $\Delta\text{-ODG} \doteq \text{ODG(CuT-D)} - \text{ODG(G.729.1-SWB)}$. All codec bit rates and test items as used in Section 6.4 have been included.

Table D.1 compares other important codec characteristics. The complexity figures for G.729.1-SWB have been measured from the available *fixed point* implementation. The complexity figures for CuT-D are given for the *floating point* implementation. The numbers in parentheses denote an *estimated fixed point complexity* with a conversion penalty factor of 1.2.

From the direct comparison of CuT-D and G.729.1-SWB (and from the respective test results in Appendix 6), it can be concluded that CuT-D ("Candidate Codec B") indeed outperforms the standardized version of the codec.

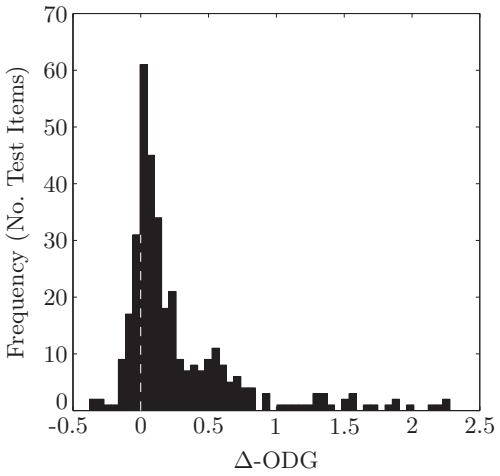


Figure D.3: Histogram of PEAQ score difference between CuT-D and G.729.1-SWB, measured over the entire EBU SQAM corpus [EBU 1988].

Table D.1: Codec characteristics of CuT-D and G.729.1-SWB.

Codec	CuT-D	ITU-T G.729.1-SWB
additional algorithmic delay [ms]	+2.21875 ms	+6.75 ms
enc. compl. (36 kbit/s) [WMOPS]	5.03 (6.04)	10.19
dec. compl. (36 kbit/s) [WMOPS]	5.27 (6.32)	5.12
enc. compl. (40 kbit/s) [WMOPS]	5.42 (6.50)	10.46
dec. compl. (40 kbit/s) [WMOPS]	5.57 (6.68)	5.16
finalized in	07/2008	03/2010

Bibliography

3GPP2 (2005), '3GPP2 TSG-C C.S0052-A: Source-controlled variable-rate multimode wideband speech codec (VMR-WB), service options 62 and 33 for spread spectrum systems'.

3GPP2 (2010), '3GPP2 TSG-C C.S0014-D: Enhanced variable rate codec, speech service options 3, 68, 70, and 73 for wideband spread spectrum digital systems'.

Adoul, J.-P. & Laflamme, C. (1997*a*), 'Depth-first algebraic-codebook search for fast coding of speech'. US Patent 5701392.

Adoul, J.-P. & Laflamme, C. (1997*b*), 'Fast sparse-algebraic-codebook search for efficient speech coding'. US Patent 5699482.

Agiomyrgiannakis, Y. & Stylianou, Y. (2004), Combined estimation/coding of highband spectral envelopes for speech spectrum expansion, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 469–472.

Agiomyrgiannakis, Y. & Stylianou, Y. (2007), 'Conditional vector quantization for speech coding', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(2), 377–386.

Andersen, S., Duric, A., Astrom, H., Hagen, R., Kleijn, W. & Linden, J. (2004), 'Internet low bit rate codec (iLBC)', IETF RFC 3951.

Aoki, N. (2009), Improvement of a band extension technique for G.711 telephony speech by using steganography, *in* 'Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)', Kyoto, Japan, pp. 487–490.

Bauer, P. & Fingscheidt, T. (2008), An HMM-based artificial bandwidth extension evaluated by cross-language training and test, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Las Vegas, NV, USA, pp. 4589–4592.

Bauer, P., Jung, M.-A., Qi, J. & Fingscheidt, T. (2010), On improving speech intelligibility in automotive hands-free systems, *in* 'Proceedings of IEEE International Symposium on Consumer Electronics (ISCE)', Braunschweig, Germany.

Bellamy, J. (1991), *Digital Telephony*, Wiley Series in Telecommunications, second edn, John Wiley & Sons Ltd., New York, NY, USA.

Bessette, B., Salami, R., Lefebvre, R., Jelínek, M., Rotola-Pukkila, J., Vainio, J., Mikkola, H. & Järvinen, K. (2002), 'The adaptive multirate wideband speech codec (AMR-WB)', *IEEE Transactions on Speech and Audio Processing* **10**(8), 620–636.

Breiman, L., Friedman, J., Olshen, R. A. & Stone, C. (1984), *Classification and Regression Trees*, CRC Press, Boca Raton, FL, USA.

Byun, K. J., Jung, H. B., Hahn, M. & Kim, K. S. (2002), A fast ACELP codebook search method, *in* 'Proceedings of 6th International Conference on Signal Processing (ICSP)', Beijing, P.R. of China, pp. 422–425.

Carl, H. & Heute, U. (1994), Bandwidth enhancement of narrow-band speech signals, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Edinburgh, Scotland, pp. 1178–1181.

Celik, M., Sharma, G. & Tekalp, M. A. (2005), Pitch and duration modification for speech watermarking, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Philadelphia, PA, USA, pp. 17–20.

Chen, B. & Wornell, G. W. (2001), 'Quantization index modulation: a class of provably good methods for digital watermarking and information embedding', *IEEE Transactions on Information Theory* **47**(4), 1423–1443.

Chen, F.-K., Chen, G.-M., Su, B.-K. & Tsai, Y.-R. (2010), 'Unified pulse-replacement search algorithms for algebra codebooks of speech coders', *IET Signal Processing* **4**(6), 658–665.

Chen, J.-H. & Gersho, A. (1995), 'Adaptive postfiltering for quality enhancement of coded speech', *IEEE Transactions on Speech and Audio Processing* **3**(1), 59–71.

Chen, O. T.-C. & Liu, C.-H. (2007), 'Content-dependent watermarking scheme in compressed speech with identifying manner and location of attacks', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(5), 1605–1616.

Chen, S. & Leung, H. (2007), Speech bandwidth extension by data hiding and phonetic classification, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 593–596.

Cheng, Q. & Sorensen, J. (2001), Spread spectrum signaling for speech watermarking, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 1337–1340.

Chennoukh, S., Gerrits, A., Miet, G. & Sluijter, R. (2001), Speech enhancement via frequency bandwidth extension using line spectral frequencies, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 665–668.

Chétry, N. & Davies, M. (2006), Embedding side information into a speech codec residual, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Florence, Italy.

Cohen, A. S., Draper, S. C., Martinian, E. & Wornell, G. W. (2006), 'Stealing bits from a quantized source', *IEEE Transactions on Information Theory* **52**(7), 2965–2985.

Cohen, G., Honkala, I., Litsyn, S. & Lobstein, A. (1997), *Covering Codes*, Elsevier. North-Holland Mathematical Library, Volume 54.

Costa, M. (1983), 'Writing on dirty paper', *IEEE Transactions on Information Theory* **29**(3), 439–441.

Cover, T. M. & Thomas, J. A. (1991), *Elements of Information Theory*, Wiley Series in Telecommunications.

Cox, I. J., Miller, M. L., Bloom, J. A., Fridrich, J. & Kalker, T. (2008), *Digital Watermarking and Steganography*, second edn, Morgan Kaufmann, Burlington, MA, USA.

Cox, I. J., Miller, M. L. & McKellips, A. L. (1999), 'Watermarking as communications with side information', *Proceedings of the IEEE* **87**(7), 1127–1141.

Daudet, L. & Sandler, M. (2004), 'MDCT analysis of sinusoids: Exact results and applications to coding artifacts reduction', *IEEE Transactions on Speech and Audio Processing* **12**(3), 302–312.

De Meuleneire, M., Taddei, H., de Zélicourt, O., Pastor, D. & Jax, P. (2006), A CELP-wavelet scalable wideband speech coder, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Toulouse, France, pp. 697–700.

Deshpande, M. M. & Ramakrishnan, K. R. (2005), A novel BWE scheme based on spectral peaks in G.729 compressed domain, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Antalya, Turkey.

Dietz, M., Liljeryd, L., Kjörling, K. & Kunz, O. (2002), Spectral band replication, a novel approach in audio coding, *in* '112th convention of the Audio Engineering Society', Munich, Germany.

Ding, H. (2004), Wideband audio over narrowband low-resolution media, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 489–492.

EBU (1988), 'European broadcast union, EBU Tech 3253: Sound quality assessment material — recordings for subjective tests'.

Eggers, J. J., Bäuml, R., Tzschoppe, R. & Girod, B. (2003), 'Scalar costa scheme for information embedding', *IEEE Transactions on Signal Processing* **51**(4), 1003–1019.

Ehara, H., Morii, T., Oshikiri, M. & Yoshida, K. (2005), Predictive VQ for bandwidth scalable LSP quantization, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Philadelphia, PA, USA, pp. 137–140.

Ehret, A., Dietz, M. & Kjörling, K. (2003), State-of-the-art audio coding for broadcasting and mobile applications, *in* '114th convention of the Audio Engineering Society', Amsterdam, The Netherlands.

Eksler, V. & Jelínek, M. (2011), Coding of unquantized spectrum sub-bands in superwideband audio codecs, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Prague , Czech Republic, pp. 5228–5231.

Ekstrand, P. (2002), Bandwidth extension of audio signals by spectral band replication, *in* 'Proceedings of IEEE Benelux workshop on MPCA', Louvain, Belgium, pp. 53–58.

Ekudden, E., Hagen, R., Johansson, I. & Svedberg, J. (1999), The adaptive multirate speech coder, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Porvoo, Finland, pp. 117–119.

Enbom, N. & Kleijn, W. B. (1999), Bandwidth expansion of speech based on vector quantization of the mel frequency cepstral coefficients, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Porvoo, Finland, pp. 171–173.

Erdmann, C. (2005), Hierarchical Vector Quantization: Theory and Application to Speech Coding, Dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany. Volume 19 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany.

Erdmann, C., Bauer, D. & Vary, P. (2002), Pyramid CELP: Embedded speech coding for packet communications, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Orlando, FL, USA, pp. 181–184.

Erdmann, C., Vary, P., Fischer, K., Xu, W., Marke, M., Fingscheidt, T., Varga, I., Kaindl, M., Quinquis, C., Kövesi, B. & Massaloux, D. (2001), A candidate proposal for a 3GPP adaptive multi-rate wideband speech codec, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 757–760.

Erez, U. & Zamir, R. (2004), 'Achieving 1/2 log (1+SNR) on the AWGN channel with lattice encoding and decoding', *IEEE Transactions on Information Theory* **50**(10), 2293–2314.

Esch, T., Heese, F., Geiser, B. & Vary, P. (2010), Wideband noise suppression supported by artificial bandwidth extension techniques, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Dallas, TX, USA, pp. 4790–4793.

Esteban, D. & Galand, C. (1977), Application of quadrature mirror filters to split band voice coding schemes, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Hartford, CT, USA, pp. 191–195.

Estrada, A. X., Bugg, M., Rabi, M. & Yip, W. (1996), Forward error correction for CELP encoded speech, *in* 'Conference Record of the 30th Asilomar Conference on Signals, Systems and Computers', Pacific Grove, CA, USA, pp. 775–778.

ETSI (1990), 'ETSI EN 300 961: Digital cellular telecommunications system (phase 2+); full rate speech; transcoding (GSM 06.10 version 8.1.1 release 1999)'. Current release: 2000.

ETSI (1998), 'ETSI EN 300 726: Digital cellular telecommunications system (phase 2+); enhanced full rate (EFR) speech transcoding (GSM 06.60 version 8.0.1 release 1999)'. Current release: 2000.

ETSI (1999), 'ETSI EN 128 062: Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); LTE; inband tandem free operation (TFO) of speech codecs; service description; stage 3 (3GPP TS 28.062 version 9.0.0 release 9)'. Current release: 2010.

ETSI (2000), 'ETSI EN 301 704: Digital cellular telecommunications system (phase 2+) (GSM); adaptive multi-rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 release 1998)'.

ETSI (2001*a*), 'ETSI TS 100 910: Digital cellular telecommunication system (phase 2+); radio transmission and reception (GSM 05.05 version 8.9.0)'.

ETSI (2001*b*), 'ETSI TS 126 190: Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); LTE; speech codec speech processing functions; adaptive multi-rate - wideband (AMR-WB) speech codec; transcoding functions (3GPP TS 26.190 version 8.0.0 release 8)'. Current release: 2009.

ETSI (2004*a*), 'ETSI TS 126 290: Digital cellular telecommunications system (phase 2+); universal mobile telecommunications system (UMTS); LTE; audio codec processing functions; extended adaptive multi-rate - wideband (AMR-WB+) codec; transcoding functions (3GPP TS 26.290 version 10.0.0 release 10)'. Current release: 2011.

ETSI (2004*b*), 'ETSI TS 126.404: Enhanced aacPlus general audio codec; encoder specification; spectral band replication (SBR) part'.

ETSI (2005), 'ETSI TS 100 909: Digital cellular telecommunication system (phase 2+); channel coding (GSM 05.03 version 8.9.0)'.

Falahati, A., Soleimani, M. & Tabataba Vakili, V. (2008), Dynamic tree pruning method for fast ACELP search, *in* 'Proceedings of 3rd International Conference Information and Communication Technologies: From Theory to Applications (ICTTA)', Damascus, Syria.

Fielder, L. D., Bosi, M., Davidson, G., Davis, M., Todd, C. & Vernon, S. (1996), AC-2 and AC-3: Low-complexity transform-based audio coding, *in* 'Audio Engineering Society Conference: Collected Papers on Digital Audio Bit-Rate Reduction', pp. 54–72.

Fischer, R. F. H. & Bäuml, R. (2004), 'Lattice costa schemes using subspace projection for digital watermarking', *European Transaction on Telecommunications* **15**(4), 351–362.

Galand, F. & Kabatiansky, G. (2003), Information hiding by coverings, *in* 'Proceedings of IEEE Information Theory Workshop (ITW)', Paris, France, pp. 151–154.

Geiser, B., Jax, P. & Vary, P. (2005), Artificial bandwidth extension of speech supported by watermark-transmitted side information, *in* 'Proceedings of INTERSPEECH', Lisbon, Portugal, pp. 1497–1500.

Geiser, B., Jax, P., Vary, P., Taddei, H., Gartner, M. & Schandl, S. (2006), A qualified ITU-T G.729EV codec candidate for hierarchical speech and audio coding, *in* 'Proceedings of IEEE Workshop on Multimedia Signal Processing (MMSP)', Victoria, BC, Canada, pp. 114–118.

Geiser, B., Jax, P., Vary, P., Taddei, H., Schandl, S., Gartner, M., Guillaumé, C. & Ragot, S. (2007a), 'Bandwidth extension for hierarchical speech and audio coding in ITU-T Rec. G.729.1', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(8), 2496–2509.

Geiser, B., Krüger, H. & Vary, P. (2010), Super-wideband bandwidth extension for wideband audio codecs using switched spectral replication and pitch synthesis, *in* 'Proceedings of German Annual Conference on Acoustics (DAGA)', Berlin, Germany, pp. 663–664.

Geiser, B., Krüger, H., Löllmann, H. W., Vary, P., Zhang, D., Wan, H., Li, H. & Zhang, L. (2009), Candidate proposal for ITU-T super-wideband speech and audio coding, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Tapei, Taiwan, pp. 4121–4124.

Geiser, B., Mertz, F. & Vary, P. (2008), Steganographic packet loss concealment for wireless VoIP, *in* 'ITG-Fachtagung Sprachkommunikation', Aachen, Germany.

Geiser, B., Ragot, S. & Taddei, H. (2008), Embedded speech coding: From G.711 to G.729.1, *in* R. Martin, U. Heute & C. Antweiler, eds, 'Advances in Digital Speech Transmission', John Wiley & Sons, Ltd., Chichester, UK, chapter 8, pp. 201–247.

Geiser, B., Roggendorf, M. & Vary, P. (2010), Multi-band pre-echo control using a filterbank equalizer, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Aalborg, Denmark, pp. 244–248.

Geiser, B., Schäfer, M. & Vary, P. (2011), Binaural wideband telephony using steganography, *in* 'Konferenz Elektronische Sprachsignalverarbeitung (ESSV)', Aachen, Germany, pp. 132–137.

Geiser, B., Taddei, H. & Vary, P. (2007), Artificial bandwidth extension without side information for ITU-T G.729.1, *in* 'Proceedings of INTERSPEECH', Antwerp, Belgium, pp. 2493–2496.

Geiser, B. & Vary, P. (2007*a*), Backwards compatible wideband telephony in mobile networks: CELP watermarking and bandwidth extension, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 533–536.

Geiser, B. & Vary, P. (2007*b*), Estimation of bandwidth extension parameters in ITU-T G.729.1, *in* 'Proceedings of ETSI Workshop on Speech and Noise in Wideband Communication', Sophia Antipolis, France.

Geiser, B. & Vary, P. (2008*a*), Beyond wideband telephony - Bandwidth extension for super-wideband speech, *in* 'Proceedings of German Annual Conference on Acoustics (DAGA)', Dresden, Germany, pp. 635–636. Special Session "Wideband Speech Revisited".

Geiser, B. & Vary, P. (2008*b*), High rate data hiding in ACELP speech codecs, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Las Vegas, NV, USA, pp. 4005–4008.

Geiser, B. & Vary, P. (2009), Joint pre-echo control and frame erasure concealment for VoIP audio codecs, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Glasgow, Scotland, pp. 1259–1263.

Gray, R. M. & Neuhoff, D. L. (1998), 'Quantization', *IEEE Transactions on Information Theory* **44**(6), 2325–2383.

Gurijala, A. (2007), Speech Watermarking through Parametric Modeling, Dissertation. Michigan State University, East Lansing, MI, USA.

Gurijala, A. & Deller, J. R. (2007), On the robustness of parametric watermarking of speech, *in* 'Proceedings of International Conference on Multimedia Content Analysis and Mining (MCAM)', Weihai, P.R. of China, pp. 501–510.

Gustafsson, H., Claesson, I. & Lindgren, U. (2001), Speech bandwidth extension, *in* 'Proceedings of IEEE International Conference on Multimedia and Expo (ICME)', Tokyo, Japan, pp. 809–812.

Hersent, O., Petit, J.-P. & Gurle, D. (2005*a*), *Beyond VoIP Protocols: Understanding Voice Technology and Networking Techniques for IP Telephony*, John Wiley & Sons Ltd., Chichester, UK.

Hersent, O., Petit, J.-P. & Gurle, D. (2005*b*), *IP Telephony: Deploying Voice-Over-IP Protocols*, John Wiley & Sons Ltd., Chichester, UK.

Hofbauer, K., Kubin, G. & Kleijn, W. B. (2009), 'Speech watermarking for analog flat-fading bandpass channels', *IEEE Transactions on Audio, Speech, and Language Processing* **17**(8), 1624–1637.

Holma, H. & Toskala, A., eds (2004), *WCDMA for UMTS*, third edn, John Wiley & Sons Ltd., Chichester, UK.

Iser, B., Minker, W. & Schmidt, G. (2008), *Bandwidth Extension of Speech Signals*, Vol. 13 of *Lecture Notes in Electrical Engineering*, Springer, Boston, MA, USA.

ISO (2005), 'ISO/IEC JTC1/SC29/WG11 MPEG 14496-3: Information technology – Coding of audio-visual objects – Part 3: Audio'.

Itakura, F. (1975), 'Line spectrum representation of linear predictor coefficients of speech signals', *Journal of the Acoustical Society of America* **57**. Supplement no. 1 (89th meeting of the Acoustical Society of America), p. 35.

Ito, A., Abe, S. & Suzuki, Y. (2009), Information hiding for G.711 speech based on substitution of least significant bits and estimation of tolerable distortion, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 1409–1412.

ITU-R (1997), 'ITU-R Rec. BS.1285: Pre-selection methods for the subjective assessment of small impairments in audio systems'.

ITU-R (1998), 'ITU-R Rec. BS.1387: Method for objective measurements of perceived audio quality'.

ITU-T (1972), 'ITU-T Rec. G.711: Pulse code modulation (PCM) of voice frequencies'. Current release: 1988.

ITU-T (1976), 'ITU-T Rec. P.48: Specification for an intermediate reference system'. Current release: 1988.

ITU-T (1984), 'ITU-T Rec. G.722: 7 kHz audio-coding within 64 kbit/s'. Current release: 1988.

ITU-T (1990), 'ITU-T Rec. G.726: 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)'.

ITU-T (1993*a*), 'ITU-T Rec. G.191: Software tools for speech and audio coding standardization'. Current release: 2010.

ITU-T (1993*b*), 'ITU-T Rec. P.56: Objective measurement of active speech level'. Current release: 2011.

ITU-T (1995), 'ITU-T Rec. P.341: Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals'. Current release: 2011.

ITU-T (1996*a*), 'ITU-T Rec. G.723.1: Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s'. Current release: 2006.

ITU-T (1996*b*), 'ITU-T Rec. G.729: Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear prediction (CS-ACELP)'. Current release: 2007.

ITU-T (1996*c*), 'ITU-T Rec. P.800: Methods for subjective determination of transmission quality'. Former Rec. P.80.

ITU-T (1999), 'ITU-T Rec. G.722.1: Low-complexity coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss'. Current release: 2005.

ITU-T (2001), 'ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs'.

ITU-T (2002), 'ITU-T Rec. G.722.2: Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband (AMR-WB)'. Current release: 2003.

ITU-T (2005), 'ITU-T Rec. P.862.2: Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs'.

ITU-T (2006), 'ITU-T Rec. G.729.1: G.729-based embedded variable bit-rate coder: An 8-32 kbit/s scalable wideband coder bitstream interoperable with G.729'.

ITU-T (2008*a*), 'ITU-T Rec. G.718: Frame error robust narrow-band and wide-band embedded variable bit-rate coding of speech and audio from 8-32 kbit/s'.

ITU-T (2008*b*), 'ITU-T Rec. G.719: Low-complexity, full-band audio coding for high-quality, conversational applications'.

Iwakiri, M. & Matsui, K. (1999), Embedding a text into conjugate structure algebraic code excited linear prediction audio codes, *in* 'Proceedings of IPSJ Computer System Symposium', Shizuoka, Japan, pp. 2623–2630. In Japanese.

Järvinen, K., Vainio, J., Kapanen, P., Honkanen, T., Haavisto, P., Salami, R., Laflamme, C. & Adoul, J.-P. (1997), GSM enhanced full rate speech codec, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Munich, Germany, pp. 771–774.

Jax, P. (2002), Enhancement of Bandlimited Speech Signals: Algorithms and Theoretical Bounds, Dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany. Volume 15 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany.

Jax, P., Geiser, B., Schandl, S., Taddei, H. & Vary, P. (2006*a*), An embedded scalable wideband codec based on the GSM EFR codec, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Toulouse, France, pp. 5–8.

Jax, P., Geiser, B., Schandl, S., Taddei, H. & Vary, P. (2006*b*), A scalable wideband "add-on" for the G.729 speech codec, *in* 'ITG-Fachtagung Sprachkommunikation', Kiel, Germany.

Jax, P. & Vary, P. (2000), Wideband extension of telephone speech using a Hidden Markov model, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Delavan, WI, USA, pp. 133–135.

Jax, P. & Vary, P. (2002), An upper bound on the quality of artificial bandwidth extension of narrowband speech signals, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Orlando, FL, USA, pp. 237–240.

Jax, P. & Vary, P. (2003), 'On artificial bandwidth extension of telephone speech', *Signal Processing* **83**(8), 1707–1719.

Jax, P. & Vary, P. (2004), Feature selection for improved bandwidth extension of speech signals, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 697–700.

Jax, P. & Vary, P. (2006), 'Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding?', *IEEE Communications Magazine* **44**(5), 106–111.

Jelínek, M., Salami, R., Ahmadi, S., Bessette, B., Gournay, P. & Laflamme, C. (2004), On the architecture of the cdma2000® variable-rate multimode wideband (VMR-WB) speech coding standard, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 281–284.

Jung, S.-K., Ragot, S., Lamblin, C. & Proust, S. (2008), An embedded variable bit-rate coder based on GSM EFR: EFR-EV, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Las Vegas, NV, USA, pp. 4765–4768.

Kay, S. M. (1993), *Fundamentals of Statistical Signal Processing — Estimation Theory*, Prentice-Hall, Inc., Upper Saddle River, New Jersey.

Kim, K. T., Choi, J.-Y. & Kang, H.-G. (2007), 'Perceptual relevance of the temporal envelope to the speech signal in the 4–7 kHz band', *Journal of the Acoustical Society of America* **122**(3), EL88–EL94.

Kim, K.-T., Lee, M.-K. & Kang, H.-G. (2008), 'Speech bandwidth extension using temporal envelope modeling', *IEEE Signal Processing Letters* **15**, 429–432.

Kondoz, A. M. (2004), *Digital Speech Coding for Low Bit Rate Communication Systems*, second edn, John Wiley & Sons Ltd., Chichester, UK.

Kontio, J., Laaksonen, L. & Alku, P. (2007), 'Neural network-based artificial bandwidth expansion of speech', *IEEE Transactions on Audio, Speech, and Language Processing* **15**(3), 873–881.

Kornagel, U. (2003), Synthetische Tiefpaß-Erweiterung von Telefonsprache, *in* 'Proceedings of German Annual Conference on Acoustics (DAGA)', Aachen, Germany, pp. 752–753. In German.

Kornagel, U. (2006), 'Techniques for artificial bandwidth extension of telephone speech', *Signal Processing* **86**(6), 1296–1306.

Kozachenko, L. F. & Leonenko, N. N. (1987), 'Sample estimate of the entropy of a random vector', *Problems of Information Transmission (Problemy Peredachi Informatsii)* **23**(2), 9–16. In Russian.

Kraskov, A., Stögbauer, H. & Grassberger, P. (2004), 'Estimating mutual information', *Physical Review E* **69 (066138)**.

Krishnan, V., Rajendran, V., Kandhadai, A. & Manjunath, S. (2007), EVRC-Wideband: The new 3GPP2 wideband vocoder standard, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 333–336.

Krüger, H., Geiser, B. & Vary, P. (2010), Gosset low complexity vector quantization with application to audio coding, *in* 'ITG-Fachtagung Sprachkommunikation', Bochum, Germany.

Krüger, H., Geiser, B., Vary, P., Li, H. T. & Zhang, D. (2011*a*), A fast indexing method for shells of the Gosset lattice, *in* 'Konferenz Elektronische Sprachsignalverarbeitung (ESSV)', Aachen, Germany, pp. 205–212.

Krüger, H., Geiser, B., Vary, P., Li, H. T. & Zhang, D. (2011*b*), Gosset lattice spherical vector quantization with low complexity, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Prague, Czech Republic, pp. 485–488.

Krüger, H., Schreiber, R., Geiser, B. & Vary, P. (2008), On logarithmic spherical vector quantization, *in* 'Proceedings of International Symposium on Information Theory and its Applications (ISITA)', Auckland, New Zealand, pp. 600–605.

Kövesi, B., Massaloux, D. & Sollaud, A. (2004), A scalable speech and audio coding scheme with continuous bitrate flexibility, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 273–276.

Kövesi, B., Ragot, S., Gartner, M. & Taddei, H. (2008), Pre-echo reduction in the ITU-T G.729.1 embedded coder, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Lausanne, Switzerland.

Laaksonen, L., Pulakka, H., Myllylä, V. & Alku, P. (2009), 'Development, evaluation and implementation of an artificial bandwidth extension method of telephone speech in mobile terminal', *IEEE Transactions on Consumer Electronics* **55**(2), 780–787.

Laaksonen, L., Tammi, M., Malenovsky, V., Vaillancourt, T., Lee, M. S., Yamanashi, T., Oshikiri, M., Lamblin, C., Kövesi, B., Miao, L., Zhang, D., Gibbs, J. & Francois, H. (2010), Superwideband extension of G.718 and G.729.1 speech codecs, *in* 'Proceedings of INTERSPEECH', Makuhari, Chiba, Japan, pp. 2382–2385.

Laaksonen, L. & Virolainen, J. (2009), Binaural artificial bandwidth extension (B-ABE) for speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 4009–4012.

Laflamme, C., Adoul, J.-P., Su, H. Y. & Morissette, S. (1990), On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Albuquerque, NM, USA, pp. 177–180.

Lamblin, C., Quinquis, C. & Usai, P. (2008), 'ITU-T G.722.1 Annex C: The first ITU-T superwideband audio coder', *IEEE Communications Magazine* **46**(10), 116–122.

Larsen, E., Aarts, R. M. & Danessis, M. (2002), Efficient high-frequency bandwidth extension of music and speech, *in* '112th Convention of the Audio Engineering Society', Munich, Germany.

Larsen, E. & Aarts, R. M., eds (2004), *Audio Bandwidth Extension - Application of Psychoacoustics, Signal Processing and Loudspeaker Design*, John Wiley and Sons, New York, NY, USA.

Lee, E. D., Lee, M. S. & Kim, D. Y. (2003), Global pulse replacement method for fixed codebook search of ACELP speech codec, *in* 'Proceedings of 2nd IASTED International Conference on Communications, Internet and Information Technology', Scottsdale, AZ, USA, pp. 372–375.

Lee, E. D., Yun, S. H., Lee, S. I. & Ahn, J. M. (2007), 'Iteration-free pulse replacement method for algebraic codebook search', *IEE Electronics Letters* **43**(1), 59–60.

Lescuyer, P. & Lucidarme, T. (2008), *Evolved Packet System (EPS): The LTE and SAE Evolution of 3G UMTS*, John Wiley & Sons Ltd., Chichester, UK.

Li, M., Jiao, Y. & Niu, X. (2008), Reversible watermarking for compressed speech, *in* 'Proceedings of 8th International Conference on Intelligent Systems Design and Applications (ISDA)', Kaohsiung, Taiwan, pp. 197–201.

Licai, H. & Shuozhong, W. (2006), Information hinding based on GSM full rate speech coding [sic], *in* 'Proceedings of International Wireless Communications, Networking and Mobile Computing Conference (WiCOM)', Wuhan, P.R. of China, pp. 1–4.

Linde, Y., Buzo, A. & Gray, R. (1980), 'An algorithm for vector quantizer design', *IEEE Transactions on Communications* **28**(1), 84–95.

Liu, C.-M., Lee, W.-C. & Hsu, H.-W. (2003), High frequency reconstruction for band-limited audio signals, *in* 'Proceedings of International Conference on Digital Audio Effects (DAFX)', London, UK, pp. 1–6.

Liu, L., Li, M., Li, Q. & Liang, Y. (2008), Perceptually transparent information hiding in G.729 bitstream, *in* 'Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)', Harbin, P.R. of China, pp. 406–409.

Löllmann, H. W. & Vary, P. (2007), 'Uniform and warped low delay filter-banks for speech enhancement', *Speech Communication* **49**(7-8), 574–587. Special Issue on Speech Enhancement.

Löllmann, H. W. & Vary, P. (2008), Design of IIR QMF banks with near-perfect reconstruction and low complexity, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Las Vegas, NV, USA, pp. 3521–3524.

Lu, Z.-M., Yan, B. & Sun, S.-H. (2005), 'Watermarking combined with CELP speech coding for authentication', *IEICE Transactions on Information and Systems* **E88-D**(2), 330–334.

Löllmann, H. W., Hildenbrand, M., Geiser, B. & Vary, P. (2009), IIR QMF-bank design for speech and audio subband coding, *in* 'Proceedings of IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)', New Paltz, NY, USA, pp. 269–272.

Makhoul, J. & Berouti, M. (1979), High-frequency regeneration in speech coding systems, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Washington, D.C., USA, pp. 428–431.

Makinen, J., Bessette, B., Bruhn, S., Ojala, P., Salami, R. & Taleb, A. (2005), AMR-WB+: A new audio coding standard for 3rd generation mobile audio services, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Philadelphia, PA, USA, pp. 1109–1112.

Malvar, H. S. (1992), *Signal Processing with Lapped Transforms*, Artech House, Norwood, MA, USA.

Massaloux, D., Trilling, R., Lamblin, C., Ragot, S., Ehara, H., Lee, M. S., Kim, D. Y. & Bessette, B. (2007), An 8-12 kbit/s embedded CELP coder interoperable with ITU-T G.729: First stage of the new G.729.1 standard, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 1105–1108.

McCree, A. (2000), A 14 kb/s wideband speech coder with a parametric high-band model, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Istanbul, Turkey, pp. 1153–1156.

McCree, A., Unno, T., Anandakumar, A., Bernard, A. & Paksoy, E. (2001), An embedded adaptive multi-rate wideband speech coder, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 761–764.

Mertz, F. (2011), Efficient Audio Communication over Heterogeneous Packet Networks with Wireless Access, Dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany. Volume 28 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany.

Miet, G., Gerrits, A. & Valiere, J. C. (2000), Low-band extension of telephone-band speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Istanbul, Turkey, pp. 1851–1854.

Mintzer, F. (1982), 'On half-band, third-band, and Nth-band FIR filters and their design', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **30**(5), 734–738.

Moon, T. K. (1996), 'The expectation-maximization algorithm', *IEEE Signal Processing Magazine* **13**(6), 47–60.

Moulin, P. & Kötter, R. (2005), 'Data-hiding codes', *Proceedings of the IEEE* **93**(12), 2083–2126.

Nagel, F. & Disch, S. (2009), A harmonic bandwidth extension method for audio codecs, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 145–148.

Nagel, F., Disch, S. & Wilde, S. (2010), A continuous modulated single sideband bandwidth extension, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Dallas, TX, USA, pp. 357–360.

Nakatoh, Y., Tsushima, M. & Norimatsu, T. (1997), Generation of broadband speech from narrowband speech using piecewise linear mapping, *in* 'Proceedings of EUROSPEECH', Rhodes, Greece, pp. 1643–1646.

Neuendorf, M., Gournay, P., Multrus, M., Lecomte, J., Bessette, B., Geiger, R., Bayer, S., Fuchs, G., Hilpert, J., Rettelbach, N., Nagel, F., Robilliard, J., Salami, R., Schuller, G., Lefebvre, R. & Grill, B. (2009a), A novel scheme for low bitrate unified speech and audio coding - MPEG RM0, *in* '126th convention of the Audio Engineering Society', Munich, Germany.

Neuendorf, M., Gournay, P., Multrus, M., Lecomte, J., Bessette, B., Geiger, R., Bayer, S., Fuchs, G., Hilpert, J., Rettelbach, N., Salami, R., Schuller, G., Lefebvre, R. & Grill, B. (2009), Unified speech and audio coding scheme for high quality at low bitrates, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 1–4.

Nilsson, M., Gustaffson, H., Andersen, S. V. & Kleijn, W. B. (2002), Gaussian mixture model based mutual information estimation between frequency bands in speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Orlando, FL, USA, pp. 525–528.

Nilsson, M. & Kleijn, W. B. (2001), Avoiding over-estimation in bandwidth extension of telephony speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 869–872.

Nishimura, A. (2009), Data hiding in pitch delay data of the adaptive multi-rate narrow-band speech codec, *in* 'Proceedings of International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)', Kyoto, Japan, pp. 483–486.

Noisex-92: Database of recording of various noises (1992), online: `http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html`.

Nour-Eldin, A. H. & Kabal, P. (2009), Combining frontend-based memory with MFCC features for bandwidth extension of narrowband speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 4001–4004.

Nour-Eldin, A. H. & Kabal, P. (2011), Memory-based approximation of the gaussian mixture model framework for bandwidth extension of narrowband speech, *in* 'Proceedings of INTERSPEECH', Florence, Italy, pp. 1185–1188.

Nour-Eldin, A. H., Shabestary, T. Z. & Kabal, P. (2006), The effect of memory inclusion on mutual information between speech frequency bands, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Toulouse, France, pp. 53–56.

NTT (1994), 'NTT advanced technology corporation: Multilingual speech database for telephonometry', online: `http://www.ntt-at.com/products_e/speech/`.

Ohm, J.-R. & Lüke, H. D. (2010), *Signalübertragung: Grundlagen der digitalen und analogen Nachrichtenübertragungssysteme*, Springer Verlag. In German.

Olsson, M., Sultana, S., Rommer, S., Frid, L. & Mulligan, C. (2009), *SAE and the Evolved Packet Core*, Academic Press, Elsevier, Oxford, UK.

Oppenheim, A. V. & Schafer, R. W. (1995), *Zeitdiskrete Signalverarbeitung*, 2nd edn, Oldenbourg Verlag, München, Wien. In German.

Orange (2010), 'Orange accelerates the roll-out of mobile HD voice in Europe.', online: `http://www.orange.com/en_EN/press/press_releases/att00014472/GroupHDrelease.pdf`.

Oshikiri, M., Ehara, H., Morii, T., Yamanashi, T., Satoh, K. & Yoshida, K. (2007), An 8-32 kbit/s scalable wideband coder extended with MDCT-based bandwidth extension on top of a 6.8 kbit/s narrowband CELP coder, *in* 'Proceedings of INTERSPEECH', Antwerp, Belgium, pp. 1701–1704.

Oshikiri, M., Ehara, H. & Yoshida, K. (2002), A Scalable Coder Designed for 10-kHz Bandwidth Speech, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Tsukuba, Ibaraki, Japan, pp. 111–113.

Oshikiri, M., Ehara, H. & Yoshida, K. (2004), Efficient Spectrum Coding for Super-Wideband Speech and its Application to 7/10/15 kHz Bandwidth Scalable Coders, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Montreal, QC, Canada, pp. 481–484.

Park, H., Choi, Y. & Lee, D. (2002), Efficient codebook search method for ACELP speech codecs, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Tsukuba, Japan, pp. 17–19.

Park, J. S., Choi, M. Y. & Kim, H. S. (2004), Low-band extension of CELP speech coder by harmonics recovery, *in* 'Proceedings of International Symposium on

Intelligent Signal Processing and Communication Systems (ISPACS)', Seoul, South Korea, pp. 147–150.

Park, K.-Y. & Kim, H. S. (2000), Narrowband to wideband conversion of speech using GMM based transformation, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Istanbul, Turkey, pp. 1843–1846.

Paulus, J. (1997), Codierung breitbandiger Sprachsignale bei niedriger Datenrate, Dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany. Volume 6 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany (in German).

Paulus, J. & Schnitzler, J. (1996), 16 kbit/s wideband speech coding based on unequal subbands, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Atlanta, GA, USA, pp. 255–258.

Pham, T. V., Schaefer, F. & Kubin, G. (2010), A novel implementation of the spectral shaping approach for artificial bandwidth extension, *in* 'Proceedings of 3rd International Conference on Communications and Electronics (ICCE)', Nha Trang, Vietnam, pp. 262–267.

Princen, J. & Bradley, A. (1986), 'Analysis/synthesis filter bank design based on time domain aliasing cancellation', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **34**(5), 1153–1161.

Pulakka, H. & Alku, P. (2011), 'Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum', *IEEE Transactions on Audio, Speech, and Language Processing* **19**(7), 2170–2183.

Pulakka, H., Laaksonen, L., Vainio, M., Pohjalainen, J. & Alku, P. (2008), 'Evaluation of an artificial speech bandwidth extension method in three languages', *IEEE Transactions on Acoustics, Speech, and Signal Processing* **16**(6), 1124–1137.

Pulakka, H., Myllylä, V., Laaksonen, L. & Alku, P. (2010), Bandwidth extension of telephone speech using a filter bank implementation for highband mel spectrum, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Aalborg, Denmark, pp. 979–983.

Pulakka, H., Rentes, U., Palomaki, K., Kurimo, M. & Alku, P. (2011), Speech bandwidth extension using gaussian mixture model-based estimation of the highband mel spectrum, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Prague , Czech Republic, pp. 5100–5103.

Ragot, S., Kövesi, B., Trilling, R., Virette, D., Duc, N., Massaloux, D., Proust, S., Geiser, B., Gartner, M., Schandl, S., Taddei, H., Gao, Y., Shlomot, E., Ehara, H., Yoshida, K., Vaillancourt, T., Salami, R., Lee, M. S. & Kim, D. Y. (2007), ITU-T G.729.1: An 8-32 kbit/s scalable coder interoperable with G.729 for wideband telephony and Voice Over IP, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 529–532.

Ramabadran, T. & Jasiuk, M. (2008), Artificial bandwidth extension of narrowband speech signals via high-band energy estimation, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Lausanne, Switzerland.

Rix, A. W., Beerends, J. G., Hollier, M. P. & Hekstra, A. P. (2001), Perceptual evaluation of speech quality (PESQ) - A new method for speech quality assessment of telephone networks and codecs, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Salt Lake City, UT, USA, pp. 749–752.

Sagi, A. & Malah, D. (2007), 'Bandwidth extension of telephone speech aided by data embedding', *EURASIP Journal on Applied Signal Processing* **2007**(1). Article 64921.

Salami, R., Laflamme, C., Adoul, J.-P., Kataoka, A., Hayashi, S., Moriya, T., Lamblin, C., Massaloux, D., Proust, S., Kroon, P. & Shoham, Y. (1998), 'Design and description of CS-ACELP: A toll quality 8 kb/s speech coder', *IEEE Transactions on Speech and Audio Processing* **6**(2), 116–130.

Sauert, B., Löllmann, H. W. & Vary, P. (2008), Near end listening enhancement by means of warped low delay filter-banks, *in* 'ITG-Fachtagung Sprachkommunikation', Aachen, Germany.

Schnell, M., Geiger, R., Schmidt, M., Jander, M., Multrus, M., Schuller, G. & Herre, J. (2007), Enhanced MPEG-4 Low Delay AAC - Low bitrate high quality communication, *in* '122nd convention of the Audio Engineering Society', Vienna, Austria.

Schnell, M., Schmidt, M., Jander, M., Multrus, M., Schuller, G. & Herre, J. (2008), MPEG-4 Enhanced Low Delay AAC - A new standard for high quality communication, *in* '125th convention of the Audio Engineering Society', San Francscio, CA, USA.

Schnitzler, J. (1998), A 13.0 kbit/s wideband speech codec based on SB-ACELP, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Seattle, WA, USA, pp. 157–160.

Schnitzler, J. (1999), Breitbandige Sprachcodierung: Zeitbereichs- und Frequenzbereichskonzepte, Dissertation, IND, RWTH Aachen University, Templergraben 55, 52056 Aachen, Germany. Volume 12 in "Aachener Beiträge zu Digitalen Nachrichtensystemen (ABDN)", Verlag Mainz, Aachen, Germany (in German).

Schroeder, M. & Atal, B. (1985), Code-excited linear prediction (CELP): High-quality speech at very low bit rates, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Tampa, FL, USA, pp. 937–940.

Schug, M., Gröschel, A., Beer, M. & Henn, F. (2003), Enhancing audio coding efficiency of MPEG Layer-2 with spectral band replication (SBR) for DigitalRadio (EUREKA 147/DAB) in a backwards compatible way, *in* '114th convention of the Audio Engineering Society', Amsterdam, The Netherlands.

Sesia, S., Toufik, I. & Baker, M., eds (2009), *LTE - The UMTS Long Term Evolution: From Theory to Practice*, John Wiley & Sons Ltd., Chichester, UK.

Shahbazi, A. et al. (2010*a*), Content dependent data hiding on GSM full rate encoded speech, *in* 'Proceedings of International Conference on Signal Acquisition and Processing (ICSAP)', Bangalore, India, pp. 68–72.

Shahbazi, A. et al. (2010*b*), MELPe coded speech hiding on enhanced full rate compressed domain, *in* 'Proceedings of Fourth Asia International Mathematical/Analytical Modelling and Computer Simulation Conference (AMS)', Kota Kinabalu, Malaysia, pp. 267–270.

Song, G.-B. & Martynovich, P. (2009), 'A study of HMM-based bandwidth extension of speech signals', *Signal Processing* **89**(10), 2036–2044.

Steele, R., Lee, C.-C. & Gould, P. (2001), *GSM, cdmaOne and 3G Systems*, John Wiley & Sons Ltd., Chichester, UK.

Takahashi, A., Kurashima, A., Morioka, C. & Yoshino, H. (2005), Objective quality assessment of wideband speech by an extension of ITU-T Recommendation P.862, *in* 'Proceedings of INTERSPEECH', Lisbon, Portugal, pp. 3153–3156.

Tammi, M., Laaksonen, L., Ramo, A. & Toukomaa, H. (2009), Scalable superwideband extension for wideband coding, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 161–164.

Taori, R., Sluijter, R. J. & Gerrits, A. J. (2000), Hi-BIN: An alternative approach to wideband speech coding, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Istanbul, Turkey, pp. 1157–1160.

Thomas, M. R. P., Gudnason, J., Naylor, P. A., Geiser, B. & Vary, P. (2010), Voice source estimation for artificial bandwidth extension of telephone speech, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Dallas, TX, USA, pp. 4794–4797.

Tian, H., Zhou, K., Jiang, H., Liu, J., Huang, Y. & Feng, D. (2009), An M-sequence based steganography model for Voice over IP, *in* 'Proceedings of IEEE International Conference on Communications (ICC)', Dresden, Germany.

Tsujino, K. & Kikuiri, K. (2009), Low-complexity bandwidth extension in MDCT domain for low-bitrate speech coding, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Taipei, Taiwan, pp. 4145–4148.

Un, C. K. & Magill, D. T. (1975), 'The residual-excited linear prediction vocoder with transmission rate below 9.6 kbit/s', *IEEE Transactions on Communications* **23**(12), 1466–1474.

Unno, T. & McCree, A. (2005), A robust narrowband to wideband extension system featuring enhanced codebook mapping, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Philadelphia, PA, USA, pp. 805–808.

Vaillancourt, T., Jelínek, M., Ertan, A. E., Stachurski, J., Rämö, A., Laaksonen, L., Gibbs, J., Mittal, U., Bruhn, S., Grancharov, V., Oshikiri, M., Ehara, H., Zhang, D., Ma, F., Virette, D. & Ragot, S. (2008), ITU-T EV-VBR: A robust 8-32 kbit/s scalable coder for error prone telecommunications channels, *in* 'Proceedings of European Signal Processing Conference (EUSIPCO)', Lausanne, Switzerland.

Vaillancourt, T., Jelínek, M., Salami, R. & Lefebvre, R. (2007), Efficient frame erasure concealment in predictive speech codecs using glottal pulse resynchronisation, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Honolulu, Hawai'i, USA, pp. 1113–1116.

Valin, J.-M. & Lefebvre, R. (2000), Bandwidth extension of narrowband speech for low bit-rate wideband coding, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Delavan, WI, USA, pp. 130–132.

Valin, J.-M., Terriberry, T. B., Montgomery, C. & Maxwell, G. (2010), 'A high-quality speech and audio codec with less than 10-ms delay', *IEEE Transactions on Audio, Speech, and Language Processing* **18**(1), 58–67.

Varga, I., Proust, S. & Taddei, H. (2009), 'ITU-T G.729.1 scalable codec for new wideband services', *IEEE Communications Magazine* **47**(10), 131–137.

Vary, P. (2006), 'An adaptive filterbank equalizer for speech enhancement', *Signal Processing* **86**(6), 1206–1214. Special Issue on Applied Speech and Audio Processing (dedicated to Prof. Hänsler).

Vary, P. & Geiser, B. (2007), Steganographic wideband telephony using narrowband speech codecs, *in* 'Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers', Pacific Grove, CA, USA, pp. 1475–1479.

Vary, P., Hellwig, K., Hofmann, R., Sluyter, R., Galand, C. & Rosso, M. (1988), Speech codec for the european mobile radio system, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', New York, NY, USA, pp. 227–230.

Vary, P. & Heute, U. (1980), 'A short-time spectrum analyzer with polyphase-network and DFT', *Signal Processing* **2**(1), 55–65.

Vary, P. & Martin, R. (2006), *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, John Wiley and Sons, Chichester, UK.

Vos, K., Jensen, S. & Sørensen, K. (2010), 'SILK speech codec', IETF Internet-Draft, Network Working Group. Work in progress. Available online: `http://tools.ietf.org/html/draft-vos-silk-01`.

Wang, T., Koishida, K., Cuperman, V., Gersho, A. & Collura, J. S. (2002), A 1200/2400 bps coding suite based on MELP, *in* 'Proceedings of IEEE Workshop on Speech Coding (SCW)', Tsukuba, Japan, pp. 90–92.

Wang, Y., Yaroslavsky, L., Vilermo, M. & Vaananen, M. (2000), Some peculiar properties of the MDCT, *in* 'Proceedings of 5th International Conference on Signal Processing (WCCC-ICSP)', Beijing, China, pp. 61–64.

Wolters, M., Kjörling, K., Homm, D. & Purnhagen, H. (2003), A closer look into MPEG-4 High Efficiency AAC, *in* '115th convention of the Audio Engineering Society', New York, NY, USA.

Xiao, B., Huang, Y. & Tang, S. (2008), An approach to information hiding in low bit-rate speech stream, *in* 'Proceedings of IEEE Global Telecommunications Conference (GLOBECOM)', New Orleans, LA, USA, pp. 1–5.

Xie, M., Chu, P., Taleb, A. & Briand, M. (2009), ITU-T G.719: A new low-complexity full-band (20 kHz) audio coding standard for high-quality conversational applications, *in* 'Proceedings of IEEE Workshop Applications of Signal Processing to Audio and Acoustics (WASPAA)', New Paltz, NY, USA, pp. 265–268.

Xie, M., Lindbergh, D. & Chu, P. (2006), ITU-T G.722.1 Annex C: A new low-complexity 14 kHz audio coding standard, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Toulouse, France, pp. 173–176.

Xu, T. & Yang, Z. (2009), Simple and effective speech steganography in G.723.1 low-rate codes, *in* 'Proceedings of International Conference on Wireless Communications & Signal Processing (WCSP)', Nanjing, P.R. of China.

Yao, S. & Chan, C.-F. (2005), Block-based bandwidth extension of narrowband speech signal by using CDHMM, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Philadelphia, PA, USA, pp. 793–796.

Yao, S. & Chan, C.-F. (2006), Speech bandwidth enhancement using state space speech dynamics, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Toulouse, France, pp. 489–492.

Yağlı C. & Erzin, E. (2011), Artificial bandwidth extension of spectral envelope with temporal clustering, *in* 'Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)', Prague , Czech Republic, pp. 5096–5099.

Zamir, R., Shamai, S. & Erez, U. (2002), 'Nested linear/lattice codes for structured multiterminal binning', *IEEE Transactions on Information Theory* **48**(6), 1250–1276.

Zhang, Q. & Boston, N. (2003), Quantization index modulation using the E8 lattice, *in* 'Proceedings of 41st Annual Allerton Conference on Communication, Control and Computing', Allerton, IL, USA.

Ziegler, T., Ehret, A., Ekstrand, P. & Lutzky, M. (2002), Enhancing MP3 with SBR. Features and capabilities of the new MP3PRO, *in* '112th Convention of the Audio Engineering Society', Munich, Germany.