Low Delay Audio Coding Based on Logarithmic Spherical Vector Quantization

Audiocodierung mit geringer Verzögerung basierend auf Logarithmisch Sphärischer Vektorquantisierung

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

Hauke Ulrich Krüger

aus Lübeck

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

Tag der mündlichen Prüfung: 24. Februar 2010

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 25)
ISSN 1437-6768
ISBN 3-86130-651-4

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, 2010)"

Acknowledgments

This thesis was written during my time as a research assistant at the *Institute of Communication Systems and Data Processing (IND)* at the *Rheinisch-Westfälische Technische Hochschule Aachen (RWTH Aachen University)*. I would like to express my gratitude to all those who contributed to the success of this work.

In particular, I am sincerely grateful to my supervisor Prof. Dr.-Ing. Peter Vary whose numerous suggestions, inspiring ideas and continuous support have been invaluable, and I highly appreciate his dedication to encourage me in my scientific and technical interest over the years. I am also indebted to the co-supervisor of my work, Prof. Dr.-Ing. Jens-Rainer Ohm, for showing much interest in the obtained results.

Furthermore, I want to thank all my colleagues, students and permanent staff at the institute for the pleasant, friendly and productive working atmosphere. For many fruitful and inspiring discussions, successful cooperative research projects and in particular the friendship, I wish to express my deepest thanks to my former colleagues Dr.-Ing. Peter Jax, Dr.-Ing. Christoph Erdmann, Dr.-Ing. Thomas Lotter, Dr.-Ing. Marc Adrat, Dr.-Ing. Thorsten Clevorn, Dipl.-Ing. Carsten Hoelper and Dr.-Ing. Gerald Enzner, to my colleagues M. Sc. Marco Jeub, Dipl.-Ing. Matthias Rüngeler, Dipl.-Ing. Bastian Sauert, Simone Sedgwick, Roswitha Fröhlich, Andreas Welbers, and the colleagues from the IND workshop. I also owe special thanks to Dipl.-Ing. Aulis Telle, Dipl.-Ing. Bernd Geiser, Dipl.-Ing. Laurent Schmalen, and Dipl.-Ing. Magnus Schäfer for proof-reading my manuscript as well as Dr.-Ing. Christiane Antweiler, Dipl.-Ing. Heiner Löllmann, and Dipl.-Ing. Thomas Schlien for valuable suggestions for improvements of the final presentations. Of course, special thanks also go to the many students who made significant contributions to my research work, in particular Dipl.-Ing. Dennis Noppeney, Dipl.-Ing. Raimund Schreiber, M. Sc. Na Zhou, and Thomas Schumacher.

I owe my loving thanks to my family, in particular, my parents Bärbel and Udo Krüger and my parents-in-law Monika and Reinhard Niggemeier for their support over the years.

Finally, I wish to thank my beloved wife Kirsten and children Jana and Moritz for their loving support, patience, and understanding.

Aachen, February 2010 *Hauke Krüger*

Abstract

Most systems for the transmission and storage of speech and audio signals are nowadays based on digital technology. For specific applications, e.g., wireless microphones for live concerts, however, operation constraints are defined which only analog technology could fulfill. The most critical and often contradictory constraints are a low algorithmic delay, a high perceived quality for speech as well as for audio signals at low bit rates and a low computational complexity. State-of-the-art standardized approaches for digital lossy source coding in general either have a high algorithmic delay or have been optimized for speech signals only and are not suitable for audio coding.

The outcome of this thesis are novel approaches for the lossy compression of digital speech and audio signals with low algorithmic delay. The new concepts are principally based on combined *linear prediction and vector quantization* which is well-known from state-of-the-art speech codecs. However, fundamental modifications of the concepts known from speech coding are essential to achieve a low algorithmic delay and a low computational complexity as well as a high perceived speech and audio quality at low bit rates.

The developed approaches for low delay audio coding significantly outperform standardized audio codecs with a comparable algorithmic delay and bit rate, e.g., the ITU-T G.722 audio codec, in terms of a higher subjective quality for speech and particularly audio signals.

Contents

Abbreviations & Mathematical Notation

List of Abbreviations

3GPP	Third Generation Partnership Program
ACELP	Algebraic Code-Excited Linear Prediction
AAC	Advanced Audio Codec
AAC-LD	Advanced Audio Codec - Low Delay
AAC-ULD	Advanced Audio Codec - Ultra Low Delay
ACF	Autocorrelation Function
ADC	Analog-to-digital Converter
ADPCM	Adaptive Differential Pulse Code Modulation
ALBVQ	Algebraic Low Bitrate Vector Quantizer
AMR	Adaptive Multirate Speechcodec
AMR-WB	Adaptive Multirate Wideband Speechcodec
APVQ	Apple Peeling Vector Quantizer
AR	Auto-regressive
CCF	Crosscorrelation Function
CD	Compact Disc
CE	Candidate Exclusion
CELP	Code-Excited Linear Prediction
DFT	Discrete Fourier Transform
DRF	Distortion Rate Function
ETSI	European Telecommunications Standards Institute
FIR	Finite Impulse Response
FM	Frequency Modulation
GIPS	Giga Instructions Per Second
GLCVQ	Gosset Low Complexity Vector Quantizer
GSM	Global System for Mobile Communications
i.i.d.	independent and identically distributed
IEC	International Electrotechnical Commission
IIR	Infinite Impulse Response
ISA	Instruction Set Architecture
ISO	International Organization for Standardization

ISDN	Integrated Services Digital Network
ITU	International Telecommunication Union
KLT	Karhunen-Loeve Transform
LMQ	Lloyd-Max Quantizer
LP	Linear Prediction
LPC	Linear Predictive Coding
LVQ	Lattice Vector Quantization
LPVQ	Linear Predictive Vector Quantization
LSVQ	Logarithmic Spherical Vector Quantization
LTP	Long Term Prediction
MC	Efficient Metric Computation
MI	Mutual Information
MIPS	Million Instructions per Second
MOS	Mean Opinion Score
MOS-LQO	(Objective) Mean Opinion Score - Listening only test
MPEG	Motion Picture Expert Group
MP3	MPEG I, audio layer III
MMSE	Minimum Mean Squared Error
MSE	Mean Squared Error
NFC	Noise Feedback Coding
PCVQ	Permutation Code Vector Quantization
PC	Personal Computer
PCM	Pulse Code Modulation
PDF	Probability Density Function
PEAQ	Perceptual Evaluation of Audio Quality
PESQ	Perceptual Evaluation of Speech Quality
PMF	Probability Mass Function
PS	Pre-selection
PSD	Power Spectral Density
RDT	Rate Distortion Theory
RDF	Rate Distortion Function
ROM	Read Only Memory
RAM	Random Access Memory
RW	Revere Waterfilling
SBC	Subband Coding
SCELP	Spherical Code-Excited Linear Prediction
SLB	Shannon Lower Bound
SNR	Signal-to-noise Ratio
SQ	Scalar Quantization
SVD	Singular Value Decomposition
SVQ	Spherical Vector Quantization
TC	Transform Coding
VQ	Vector Quantization
W-CELP	Warped Code-Excited Linear Prediction

W-SCELP	Warped Spherical Code-Excited Linear Prediction
WB-PESQ	Wideband Perceptual Evaluation of Speech Quality
WLP	Warped Linear Prediction
WLPC	Warped Linear Predictive Coding
WMOPS	Weighted Million Operations per Second

Mathematical Notation

In this thesis, the following conventions are used to denote quantities: capital bold letters refer to matrices, e.g., $\mathbf{A}$, bold letters refer to vectors, e.g., $\mathbf{a}$, and scalars are not bold, e.g., a.
Furthermore, estimated quantities are labeled with a hat, e.g., $\hat{\mathbf{y}}$, quantized variables are marked by a tilde, e.g., $\tilde{\mathbf{y}}$, and mean values are labeled by a bar, e.g., $\bar{\mathbf{y}}$.
Components in vectors are written in vertical notation, e.g.,

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_{Lv-1} \end{bmatrix}. \tag{0.1}$$

Nevertheless, in order to save space, very often, (0.1) is written as

$$\mathbf{x} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{Lv-1} \end{bmatrix}^T. \tag{0.2}$$

In order to simplify the notation, logarithmic values given in dB are often followed by the extension $|_{\mathrm{dB}}$, e.g., $G_p|_{\mathrm{dB}} := 10 \cdot \log_{10}(G_p)$.

List of Principal Symbols

Principal Symbols related to Chapter 2:

$d(\mathbf{x}, \tilde{\mathbf{x}})$	Average quantization distortion given an event of a set of continuous random variables
$d(x_i, \tilde{x}_i)$	Per coordinate quantization distortion function
D	Expectation of quantization cost function (quantization distortion)
$\mathcal{D}(R)$	Normalized distortion rate function
D_0	Common distortion for random variables in reverse waterfilling
$I(\mathbf{X}; \tilde{\mathbf{X}})$	Mutual information
$\mathcal{N}(x, \mu, \sigma^2)$	PDF for Gaussian random variable
$p_{\mathbf{X}}(\mathbf{x})$	Multivariate PDF related to set of continuous random variables

$p_(\mathbf{x})$	Multivariate PDF (abbreviation for $p_{\mathbf{X}}(\mathbf{x})$)
$p_{\tilde{\mathbf{X}}}(\mathbf{x})$	Multivariate PMF related to set of discrete random variables $\mathbf{X}$
$p_Q(\tilde{\mathbf{x}} \mid \mathbf{x})$	Conditional PDF to describe a quantizer
$p_x(x)$	PDF related to signal $x(k)$
$q(\tilde{\mathbf{x}})$	Multivariate PMF (abbreviation for $p_{\tilde{\mathbf{X}}}(\mathbf{x})$)
$R_G(D)$	Rate distortion function for a source with Gaussian distribution
$R_{\mathrm{SLB}}(D)$	The Shannon Lower Bound
$R_{\mathrm{vr}}(k)$	Instantaneous bit rate in constrained entropy scalar quantization
$\bar{R}_{\mathrm{vr}}$	Average bit rate in constrained entropy scalar quantization
$\mathbf{X}$	Set of continuous random variables
$\tilde{\mathbf{X}}$	Set of discrete random variables
X_i	Single continuous random variable
$\tilde{X}_i$	single discrete random variable
$\mathcal{X}$	Infinite alphabet for set of continuous random variables
$\tilde{\mathcal{X}}$	Finite alphabet for set of discrete random variables
$\mathbf{x}_{e_i}$	Eigenvector
σ_i^2	Variance of the i-th component in a set of random variables
$\Delta_{D,\mathbf{X}}$	Distortion rate function for Gaussian source in relation to the SLB for arbitrary PDF
φ_{X_i,X_j}	Crosscorrelation function related to random variables X_i and X_j.
$\mathbf{\Phi_X}$	Covariance matrix
$\lambda_{\mathbf{X},i}$	Eigenvalue
$\phi_{\mathbf{X}}(\omega)$	Power spectral density
$\mathbf{\Lambda_X}$	Matrix composed of Eigenvalues
σ^2	Variance

Principal Symbols related to Chapter 3:

A	A-value (constant) in A-Law quantization
$C_{\mathrm{sq},i}$	Quantization interval in scalar quantization
$F(L_v)$	Space filling vector quantization advantage
k	Time index

L_v	Dimension of a vector quantizer
$M(L_v)$	Memory vector quantization advantage
N_Q	Number of quantization reconstruction levels
N_{Q,L_v}	Number of codevectors of dimension L_v
$\mathrm{NI}(P)$	Normalized inertia in high rate vector quantization
$\mathcal{P}_{L_v}$	Amount of admissible polytopes of dimension L_v
Q	Quantizer
$S_U(L_v)$	Shape vector quantization advantage for a multivariate uniform signal distribution
$S_G(L_v)$	Shape vector quantization advantage for a multivariate Gaussian signal distribution
$S(L_v)$	Shape vector quantization advantage
$\mathrm{SNR}_{\mathrm{sq,nu}}(R_{\mathrm{fr}})$	SNR related to non uniform scalar quantization
$\mathrm{SNR}(R)$	Rate distortion function expressed as a signal-to-noise ratio
$\mathrm{SNR}_{\mathrm{sq,u}}$	SNR related to uniform scalar quantization
$\mathrm{SNR}_{\mathrm{sq,nu,A}}$	SNR related to A-Law logarithmic scalar quantization
R_{fr}	(Constant) bit rate in fixed rate scalar quantization
$V(P)$	Volume of polytope P
$x_{b,i}$	Scalar quantization interval bounds
$\tilde{x}_i$	Scalar quantization amplitude reconstruction level
Δ_{SNR}	(Asymptotic) difference between rate distortion function for correlated and uncorrelated random variables
Δ_{u}	Quantization interval size in uniform scalar quantization
$\lambda(x)$	Quantizer point density function
$\Delta_{\mathrm{SNR,nu,G}}$	SNR offset in comparison to rate distortion function related to uniform scalar quantization
$\Delta_{\mathrm{SNR,nu}}$	SNR offset in comparison to rate distortion function related to optimal non uniform scalar quantization
$\Delta_{\mathrm{SNR,u}}$	SNR offset in comparison to rate distortion function related to uniform scalar quantization for Gaussian sources
$\Delta_{\mathrm{SNR,vr}}$	SNR offset in comparison to rate distortion function related to constrained entropy uniform scalar quantization
$x(k)$	Signal in time domain
$\mathbf{x}$	Vector

Principal Symbols related to Chapter 4:

$C_{\tilde{\mathbf{c}}}$	Spherical quantization cell in LSVQ
$C_{\tilde{\mathbf{x}}}$	Overall quantization cell in LSVQ
$C_{\tilde{g}}$	Quantization interval of the (scalar) quantizer of the gain factor in LSVQ
$\mathbf{c}$	Normalized vector to be quantized
$\tilde{\mathbf{c}}$	Normalized vector to be quantized
$D^{(I)}_{\text{lsvq}}$	Distortion related to LSVQ in the qualitative (high rate) analysis
$D^{*(II)}_{\text{svq}}$	Per vector distortion derived for an "idealized" SVQ
$D^{*(II)}_{\text{lsvq}}$	Per vector distortion derived for an "idealized" LSVQ
$D^{*(III)}_{\text{svq}}$	Per vector distortion related to "idealized" SVQ (high bit rate approximations)
$D^{*(III)}_{\text{lsvq}}$	Per vector distortion related to "idealized" LSVQ (high bit rate approximations)
$D(x, Q(x))$	Quantization distortion in scalar quantization
E_8	The Gosset Lattice
E_{Lv}	The generalized Gosset Lattice
g	Gain factor in LSVQ
$\mathcal{I}_{\text{pos}}$	Allowed pulse positions in ALBVQ
N_{lsvq}	Number of LSVQ codevectors
N_{svq}	number of SVQ codevectors
N_g	number of quantization reconstruction levels of the quantizer for the gain factor in LSVQ
$N_{(A)}$	GLCVQ code construction parameter
$N_{(B)}$	ALBVQ code construction parameter
$N_{(C)}$	APVQ code construction parameter
Q_g	Scalar quantizer for gain factor in LSVQ
Q_{svq}	Spherical vector quantizer for the shape vector in LSVQ
Q_{lsvq}	Overall quantizer in LSVQ
$R_{\text{eff,lsvq}}$	Effective bit rate per vector coordinate in LSVQ
$R_{\text{eff,svq}}$	Effective bit rate per vector coordinate in SVQ
$R_{\text{eff},g}$	Effective bit rate per vector coordinate for quantization of gain factor in LSVQ

$S_{C_{\tilde{\mathbf{c}}}}^{(I)}$	Content of spherical quantization cells (high rate assumption)
$S_{C_{\tilde{\mathbf{c}}}}^{(II)}$	Content of spherical quantization cells ("idealized SVQ")
$\mathrm{SNR}_{\mathrm{lsvq}}^{(I)}$	SNR related to LSVQ in the qualitative (high rate) analysis
$\mathrm{SNR}_{\mathrm{svq}}^{(II)}$	SNR derived for an "idealized" SVQ
$\mathrm{SNR}_{\mathrm{lsvq}}^{(II)}$	SNR derived for an "idealized" LSVQ
$\mathrm{SNR}_{\mathrm{lsvq}}^{(III)}$	SNR derived for an "idealized" LSVQ for high rate approximations
$\mathcal{S}_{N_{(A)}}^{(E_{L_v})}$	Shell of the generalized Gosset Lattice (GLCVQ)
$\mathcal{S}_{N_{(B)}}^{(\mathcal{A}_{L_v})}$	ALBVQ codevectors before normalization
$\mathcal{S}_{(m_{\vartheta_1})}^{*}$	Circles on the surface of the sphere related to the APVQ codevector construction
$\mathcal{S}_{L_v}^{(r)}$	Unit sphere with radius r
$S_{\mathcal{S}_{L_v}}^{(}r)$	Area content of sphere surface
$V_{\mathcal{S}_{L_v}}^{(r)}$	Volume of sphere
$V_{C_{\tilde{\mathbf{x}}}}(\tilde{g})$	Volume of the overall quantization cell in LSVQ
$\beta_{\max}$	Maximum angular radius to define a spherical cap quantization cell in SVQ
$\Delta_g(\tilde{g})$	Size of quantization interval related to scalar quantization of the gain factor in LSVQ
$\lambda_G(\mathbf{x})$	Optimal codevector density for multivariate Gaussian PDF
δ	Sphere packing density constant in vector quantization
θ	Sphere covering thickness constant in vector quantization
$\lambda_{L_v,\mathrm{sp}}(r)$	Sphere density function in LSVQ
$\vartheta_{\mathbf{c},\nu}$	Angle in polar coordinates
δ_ν	Distance between codevectors related to the νth angle in polar coordinates

Principal Symbols related to Chapter 5:

$\mathbf{a}_{\mathrm{ar},1}$	Set of filter coefficients to control exemplary AR processes
$A(z)$	System function of linear prediction signal estimator
C_Δ	Constant for the determination of error weighting filter $F_{\mathrm{new}}(z)$
$\mathbf{e}_{i_{\mathrm{vq}}}$	Error vector in CELP encoder

$\mathbf{e}_{w,i_{\mathrm{vq}}}$	Weighted error vector in CELP encoder
$F(z)$	Error weighting filter in generalized closed-loop LPC
$F_{\mathrm{new}}(z)$	New error weighting filter
$F_{\mathrm{conv}}(z)$	Conventional error weighting filter
G_p	Maximum achievable prediction gain
$h(\mathbf{X})$	Differential entropy
$h_W(k)$	(Truncated) impulse response of combined weighting filter $H_W(z)$
$\mathbf{h}_W$	(Truncated) impulse response of combined weighting filter $H_W(z)$ in vector notation
$H_W(z)$	Combined weighting filter
$H_{W,\mathcal{S}_{A/B}}(z)$	Part of the combined weighting filter related to states $\mathcal{S}_{A/B}$
$\mathcal{M}_{i_{\mathrm{vq}}}^{S_{\mathrm{x}}}$	Quantization cost or metric in the CELP index iteration procedure
N_{ce}	Number of candidates in pool related to the Candidate Exclusion
N_{lpc}	Linear prediction order
$q_{S_{\mathrm{x}}}$	Quantization performance loss due to search strategy S_{x} with reduced complexity
$W(z)$	Error weighting filter in CELP coding (equivalent to $F(z)$)
$\mathbf{x}_{\mathrm{fr}}$	Filter ringing signal in the modified CELP approach
$\Delta(k)$	Quantization error signal amplitude
γ	Parameter for the setup of the error weighting filter (conventional approach)
$\mathbf{\Delta}$	Quantization error vector to update the states in modified CELP approach
$\Xi_{\mathrm{SF}}(x(k))$	Spectral flatness measure
$\mathbf{\Phi}_{\Delta}$	Covariance function of the quantization error

Principal Symbols related to Chapter 6:

$\mathrm{A}^w(z)$	Frequency warping allpass
$H_S^w(z)$	Warped linear prediction synthesis filter
λ^w	Frequency warping coefficient

1

Introduction

In the past decades, a "digital revolution" in communications and multimedia technologies took place:

- Wired telephony was still based on analog networks and terminals in the late 1980's when the *Integrated Services Digital Network* (ISDN) was standardized (e.g. [ITU93]) to substitute the wide spread analog technology.
- Mobile terminals were based on analog technology when the digital wireless telephony (the *D-Netz*) was introduced in Europe in 1992. In Germany, the switch from analog to digital technology was completed in 2001 with the complete deactivation of the analog *C-Netz*.
- In the early 1980's, audio signals were recorded on analog magnetic audio tapes when the *compact disc* (CD) was launched to enable storage of audio contents in a digital way. Meanwhile, the CD and various successor technologies have replaced analog recording medias almost everywhere.

In this context, the invention of technologies to represent signals in a very compact way in the digital domain based on efficient source coding algorithms has played a fundamental role. The best known and commercially successful audio coding application is the *MPEG-1, audio layer 3* audio codec, denoted as *MP3* [ISO93], which was the basis for the development of a completely new market of internet based music distribution and lifestyle [Bla04].
Even though this "digital revolution" may seem to be complete, indeed, certain applications involving speech and audio are still based on analog technology since constraints must be fulfilled which are not or only inadequately compatible with existing digital source coding techniques. Especially low algorithmic delay, high transmission robustness and high quality at low bit rates for all types of signals are barely achievable simultaneously by state-of-the-art standardized speech and audio codecs.
For this reason, a novel approach for speech and audio coding in the digital domain is developed in this thesis that enables low algorithmic delay, achieves high quality

for speech and audio signals at moderate bit rates, and which can be optimized to be robust against bit errors in wireless transmission scenarios. Special focus is drawn to theoretical analyses of the underlying principles as well as practical concepts for source coding of signals with low computational complexity and memory consumption.

1.1 Application Examples

Purely analog technology for speech and audio transmission is currently often used for applications in which a signal is available at a sender and must be brought to a receiver on two different transmission channels in parallel, e.g., via

- direct acoustic transmission and
- a transmission involving a transformation of the signal, a wireless radio link and the signal reconstruction at the receiver.

A low transmission delay is crucial to guarantee that the two transmitted signals are (at least approximately) synchronous at the receiver side. Typical application scenarios are explained in the following.

1.1.1 Wireless Microphones in Live Concerts

The application of wireless microphones in live concerts is demonstrated by Figure 1.1. A live ensemble is playing music based on drums, a guitar, and a singer for an audience in a large concert hall. The singer's voice and the guitar are recorded by a microphone, transmitted to a receiver, amplified, and finally played back through the loudspeakers since both do not produce sufficient loudness levels. In order to allow the singer and the guitar player to freely move on the stage, both microphones are linked to the receiver and amplifier via a wireless transmission link. The acoustic volume produced by the drums is high enough so that an amplification may not be necessary. Consequently, sound signals reach the audience via two different paths, the percussion sounds via the direct acoustic transmission path and the guitar and voice signals via the wireless transmission link. Low transmission delay from the microphone to the amplifier is required to combat the following problems:

- The amplified voice signal reaches the singer via the delayed microphone-to-amplifier wireless transmission link. The singer is disturbed by his own voice if the amplified microphone signal is returned with too much delay (Scenario labeled by marker "1" in the Figure).
- The music quality is degraded if the music components produced by the drums and those from the guitar/singer reach the audience asynchronously. A very low delay is not noticeable and can be tolerated (Scenario labeled by marker "2" in the Figure).

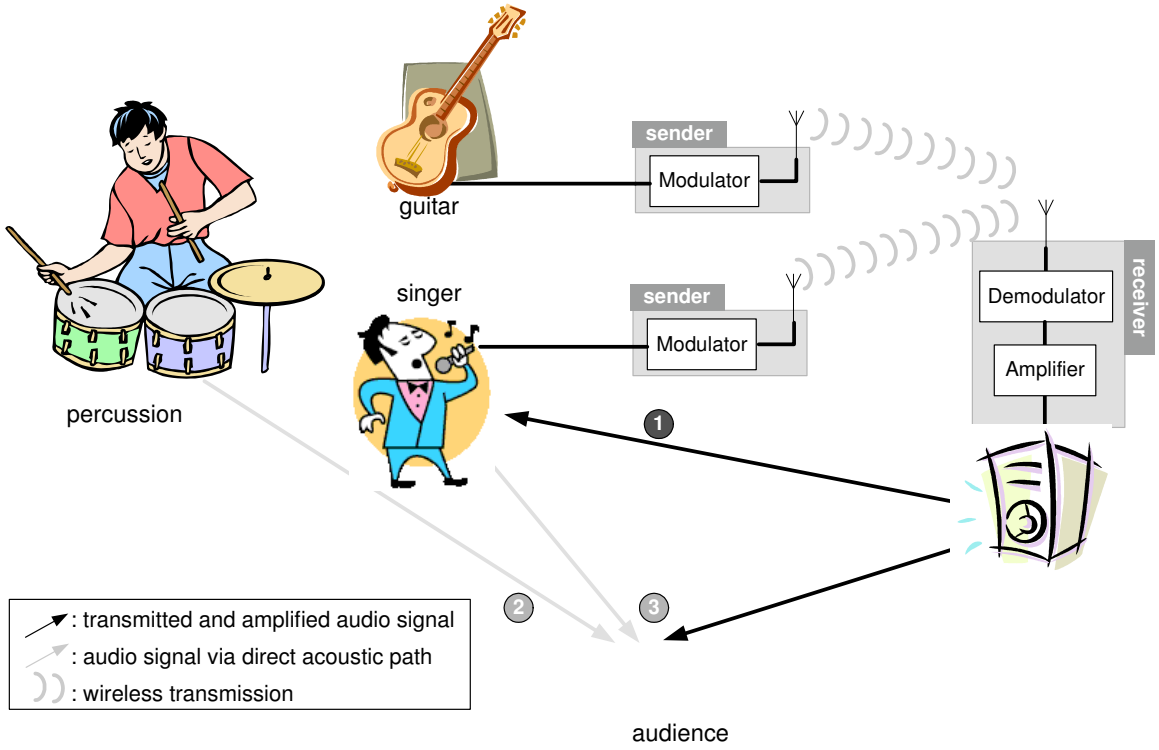


Figure 1.1: Application example for state-of-the-art wireless microphones in a live concert based on analog radio techniques (modulator and demodulator).

- The voice and the guitar signals (partly) reach the audience twice, once through direct acoustic emission and once via the microphone-to-amplifier wireless transmission link. If the delay difference between both signals is too high, undesired delay or comb filter effects result (Scenario labeled by marker "3" in the Figure).

Informal listening tests showed that a transmission delay of **less than 10 ms** can approximately be tolerated. For this reason, most wireless microphones currently available in the market are based on analog transmission systems (Frequency Modulation (FM)) or high bit rate Pulse Code Modulation (PCM) based digital transmission systems. A proper setup of a larger number of analog wireless microphones for very big concert events is often problematic due to typical analog problems such as intermodulation artifacts. Digital PCM wireless microphone systems require a high bit rate and therefore a high transmission bandwidth which limits the number of microphones.

1.1.2 Wireless Audio-Link for Hearing Aids

More and more functionality has been integrated into digital hearing aids recently so that, besides their original functionality, these devices have become sophisticated multimedia and communication units [Pud08]. In that context, modern hearing aids can be connected to other multimedia and communication devices via a wireless audio-link. This technology is exemplified in Figure 1.2 where a person wearing

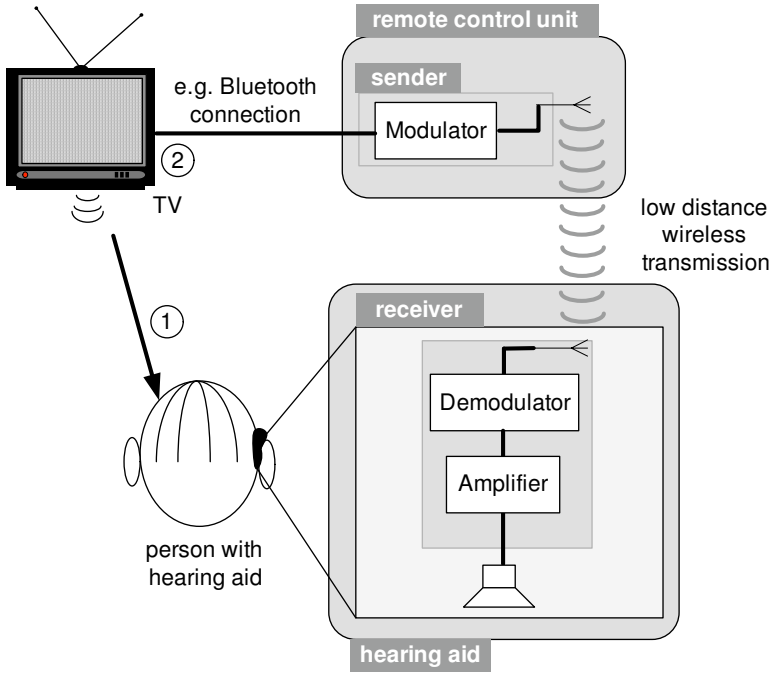


Figure 1.2: Application example for wireless audio-link in hearing aids based on analog transmission technique.

a hearing aid with integrated wireless audio-link is connected to a TV set. Since common digital transmission techniques such as bluetooth can not be directly integrated into hearing aids due to low power consumption constraints, the audio signal is transmitted to a remote control unit (RCU) first. In the next step, the signal is then forwarded from the RCU to the hearing aid via an analog low-distance wireless transmission link. Audio signals may reach the person wearing the hearing aid on a direct acoustic path and a path involving the wireless transmission link. Therefore, a low transmission delay is important (Scenario labeled by marker "1"). In addition to that, a low transmission delay is crucial to ensure synchrony between the visual content from the TV and the audio signal reaching the hearing aid (Scenario labeled by marker "2"). This becomes even more important since two different transmission techniques (Bluetooth and the low distance wireless audio-link) may be combined due to the signal routing through the RCU.

In both examples, replacing the analog transmission by digital technology yields a significantly higher transmission error robustness to overcome problems related to analog radio technology. The transmitted audio bandwidth of a digital wireless transmission system should be high, and the algorithmic delay of the involved digital source coder must be low enough not to introduce a significant amount of additional delay compared to the analog transmission system. Compared to the mentioned PCM based digital systems, the bit rate should be as low as possible so that the (radio) transmission bandwidth is narrow even in case of the operation of multiple wireless transmission units in parallel. Since the mobile devices are commonly battery powered, low computational complexity of the encoder and/or the decoder

is an additional design target for the source coder in the digital wireless transmission system.

1.2 Available Speech and Audio Codecs

Research on digital source coding for speech and audio signals was pushed by disjoint groups in the past due to completely different application constraints. Consequently, two different codec families have evolved which will be briefly described in the following.

1.2.1 Speech Coding

The development of candidates for speech coding was mostly driven by standardizations in wireless communications for narrowband (300-3400 Hz audio bandwidth) and wideband (50-7000 Hz audio bandwidth) speech. Aspects such as transmission robustness, minimal transmission bandwidth (low bit rate), and low computational complexity to enable operation in battery powered handheld mobile terminals have been important application constraints. In addition, limitations for the allowed algorithmic delay are in general specified to ensure a high conversational quality.
The latest and widest spread speech coding standards, e.g., the *Enhanced Full Rate* (EFR, [ETS96]), the *Adaptive Multi-rate* (AMR, [ETS00] and the *Adaptive Multi-rate Wideband* codec (AMR-WB, [ETS01]), are based on a model for human speech production. In order to fulfill the low delay constraints, the employed technique is time domain based and, in particular, employs linear prediction (LP) combined with vector quantization (VQ) following the *Code-Excited Linear Prediction* (CELP) [SA85] approach.
Speech codecs have been standardized mainly by the *International Telecommunication Union* (ITU) and the *European Telecommunication Standards Institute* (ETSI) or *3rd Generation Partnership Project* (3GPP) in the past. In general, latest standardized low bit rate speech codecs have an algorithmic delay of round about 20 ms and offer reasonable quality for speech signals at very low bit rates. Only specific speech coding candidates such as the *ITU-T G.711* [ITU88a], the *ITU-T G.728* [ITU92], and the *ITU-T G.722* [ITU88b] codec enable an algorithmic delay below 10 ms. However, these speech codecs mostly produce poor quality for audio signals due to the assumption of a speech production model, especially the candidates operated at very low bit rates.

1.2.2 Audio Coding

Most standardized audio codecs have been developed for music archival storage. In general, the most important aspect in audio coding is to guarantee a high or transparent quality. Constraints regarding the involved computational complexity have not been significant limitations since the audio encoder can be operated offline on, e.g., a personal computer (PC) which is often connected to the public electricity

network. Over the last years, manufacturers of mobile devices have started to integrate audio decoders on mobile battery powered platforms so that the development of low complexity decoders has been driven by market requirements. Due to the offline nature of the encoding procedure, transmission robustness and low algorithmic delay were not important issues.
In audio coding, transform coding [ZN77] based on spectral transforms such as the *Modified Discrete Cosine Transform* (MDCT, [PB86]) or the *Fast Fourier Transform* (FFT), e.g., [OS92] is mostly employed in combination with perceptual models to exploit the properties of the human auditory system. In order to achieve a sufficient spectral resolution, the transform lengths are very large so that most audio codecs have a high algorithmic delay. As the bit rate needs not to be fixed, approaches for variable bit rate coding are employed to reduce the overall average bit rate.
Audio codecs [Bra06] have been standardized mostly by the *International Organization for Standardizations* (ISO) which is often referred to as the *Moving Picture Experts Group* (MPEG). The most famous standardized audio codecs are the *MP3* codec [ISO93] and the *Advanced Audio Codec* [ISO97]. A low delay variant of the *Advanced Audio Codec* (AAC-LD) [ISO05] is available and adapts the audio coding concepts for communication systems. Audio codecs offer a high quality for audio signals at high bit rates but often are inferior compared to speech codecs for speech signals at very low bit rates.

1.2.3 Converged Speech and Audio Coding

Recently, manufacturers push **converged services** in telecommunications and multimedia networks. On the one hand, this keyword represents the demand to develop new techniques for interoperability in heterogeneous networks. On the other hand, for converged services, codecs are desired which perform well for speech and audio signals, denoted as **speech and audio convergence coding**. For this reason, novel approaches for converged speech and audio coding have been proposed in the literature, e.g., [KKO08], [BGK$^+$08]. Also new coding standards have been created to combine speech and audio coding in a hierarchical concept such as the *Extended Adaptive Multi-Rate Wideband Codec* (AMR-WB$^+$, [ETS05]) and the *ITU-T G.729.1* [ITU06] codec.

A selection of state-of-the-art standardized speech and audio codecs for higher audio bandwidths (wideband or higher) are depicted in Figure 1.3. The overall bit rate in kbit/sec is assigned to the x-axis, and the corresponding algorithmic delay to the y-axis. In the area of very low algorithmic delay (less than 10 ms), only the *ITU-T G.722* audio codec is currently freely available. Besides the standardized codecs, proprietary source coding solutions for low delay audio coding exist, e.g. the *AAC Ultra Low Delay* (AAC-ULD) codec developed by the *Fraunhofer Institute for Digital Media Technology* [KSW$^+$04].
Considering the achievable audio quality, to allow for a **high** algorithmic delay is fundamental to reach high or even transparent quality at low bit rates (e.g., the

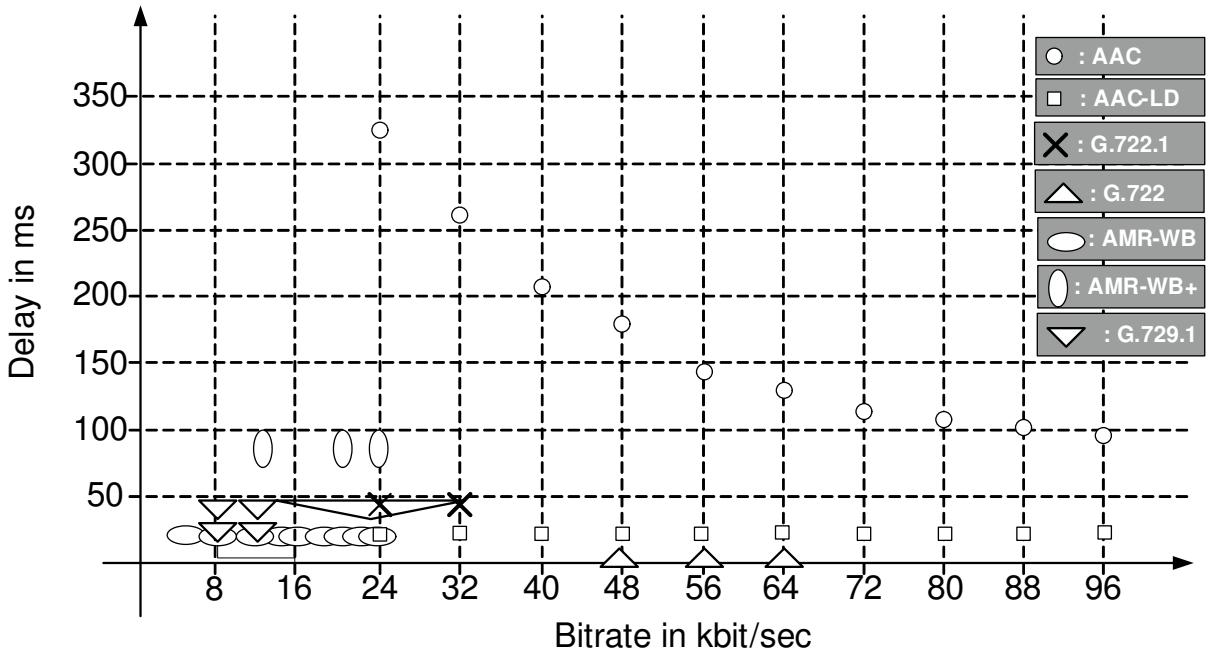


Figure 1.3: State-of-the-art standardized speech and audio codecs and corresponding bit rates and algorithmic delay.

24 kbit/sec mode of the AAC codec). Therefore, it would be unrealistic to expect that a comparable audio quality can be achieved by approaches with very **low** algorithmic delay.

1.3 The New Low Delay Speech and Audio Codec

Obviously, suitable candidates for source coding of speech and audio signals with low bit rates which have a delay below 10 ms to fulfill the constraints as defined by the example application scenarios are currently only rarely available. Those standardized codecs which have a low algorithmic delay, e.g., the ITU-T G.722 codec at data rates of 48, 56, and 64 kbit/sec, produce poor quality, especially for audio signals.
For this reason, new techniques for low delay speech and audio coding are presented in this thesis. The rough idea of the underlying concept is the following: In order to reach the desired low algorithmic delay, the time domain concept as known from speech coding is followed. Same quality for speech and audio signals is achieved by not allowing any components related to the specific characteristics of speech, e.g., no exploitation of long-term prediction to model the speakers instantaneous pitch period as commonly used in speech coding. In return, the bit rate is slightly increased by employing a new type of vector quantization to achieve a higher subjective quality for all types of signals. Correlation within the signal to be encoded is exploited by means of combined linear prediction and quantization. In that context, linear prediction must be viewed as a compact representation of the signals

spectral envelopes rather than a method to identify vocal tract parameters as often motivated in speech coding, e.g., [VM06].

1.4 Structure of the Thesis

In this thesis, general theoretical analyses and practical concepts for low delay audio coding are presented along with a concrete proposal of two new low delay audio codecs. The remainder of this thesis is divided into six chapters. In the first two chapters, theoretical results from rate distortion theory and high rate (asymptotic) quantization theory are briefly summarized: Rate distortion theory is reviewed in Chapter 2 and provides general bounds on the maximum achievable performance in quantization. In particular the *reverse waterfilling* procedure for rate distortion optimal quantization of correlated sources is described and will be of high importance in later chapters. The theory of quantization with fixed and variable rate scalar and fixed rate vector quantizers is discussed in Chapter 3. It is concluded that variable bit rate concepts are not suitable for the target applications and fixed rate vector quantization is the most promising approach that clearly outperforms fixed rate scalar quantization.
In Chapter 4, the concept of Logarithmic Spherical Vector Quantization (LSVQ), a specific type of gain-shape vector quantizer, is introduced. LSVQ is the direct consequence of the application of the high rate asymptotic vector quantization theory to approximate the optimal normalized codevector density for the Gaussian distribution as a worst case assumption for the distribution of samples in the context of audio signals. It will be shown qualitatively that the signal-to-noise ratio (SNR) achieved by LSVQ is approximately independent from the input signal distribution. A detailed theoretical analysis yields novel (quantitative) lower bounds for the achievable quantization distortion. The theoretical analysis of LSVQ is followed by the proposal of three different practical LSVQ concepts. Due to the development of novel approaches for nearest neighbor quantization, these concepts can be realized with very low computational complexity and memory consumption to be well applicable in practical applications. The chapter concludes with a comparison of the SNR measured for the three proposed LSVQ concepts and the theoretical results.
In Chapter 5, LSVQ is combined with linear prediction. Even though combined linear prediction and quantization is principally well-known from speech coding, new aspects are investigated which are of high relevance for the coding of audio signals with low algorithmic delay. In particular, novel theoretical results are derived which, in contrast to the high rate theory of linear predictive coding well-known from the literature, are valid also for lower bit rates. A conclusion drawn from the new theory is that linear predictive quantization with feedback of the quantization error can become instable. A new optimization criterion for the block adaptive computation of the filter coefficients involved in closed-loop linear predictive coding is derived which is shown to be the approximation of the reverse waterfilling

principle known from rate distortion theory.
In Chapter 6, the concepts and theoretical investigations from the previous chapters form the basis for the derivation of two new and very flexible low delay audio codecs, denoted as the Spherical Code-Excited Linear Prediction (SCELP) and the Warped Spherical Code-Excited Linear Prediction (W-SCELP) codec. In order to develop the SCELP codec, the concept of combined LSVQ and linear prediction is optimized in terms of the computational complexity. Based on measurements of signal-to-noise ratios, it is shown that by exploiting the properties of one specific among the proposed LSVQ concepts a huge reduction of computational complexity can be achieved while the quality is only marginally decreased. The W-SCELP codec is the extension of the SCELP codec and employs frequency warped linear prediction (WLP) to account for the properties of the human auditory system. Therefore it achieves a higher perceived quality especially for audio signals.
Measurements of the computational complexity of fixed point implementations of the proposed codecs show that both codecs can be operated with moderate complexity. Quality assessments with objective quality measures document that the new codecs significantly outperform standardized codecs with a comparable delay and bit rate, e.g., the ITU-T G.722 codec, in terms of a higher subjective quality for speech and particularly audio signals.

Parts of the results of this thesis are presented in the following references published by the author: [GKL+09, KGV08, KJLV09, KLEV03, Krü09, KSE+09, KSGV08, KV02, KSV06, KV05, KV06a, KV06b, KV07a, KV07b, KV08a, KV08b, KV08c, SKV09]. These references are highlighted by underlines in the following, i.e., [____]. Note that this thesis is accompanied by a *supplement document* [Krü09] to provide interested readers with additional and very detailed information on selected topics.

2

Rate Distortion Theory

A vast number of source coding algorithms targeting the lossy compression of digital signals have been proposed in the literature. The goal of all these algorithms is to achieve a high bit rate reduction while retaining a high perceptual quality. Always, the choice of the employed technology is highly influenced by application specific constraints such as, e.g., type of source signal, allowed coding delay, or available data rate. Even though in some source coding algorithms well hidden behind a lot of pre- and postprocessing, the common key component with a direct impact on bit rate and quality in lossy source coding is the quantizer. A proper design is very important as a suboptimally designed quantizer is a burden that can never be compensated by any additional signal processing.
In the design of a new quantizer, it is very useful to know the maximum theoretically achievable quantization performance. A very general approach to provide this information was given by the "Mathematical Theory of Communication" [Sha48] and, specifically, the *Rate Distortion Theory* (RDT).
The RDT provides bounds which will be presented in this section. In the remainder of this thesis, these bounds will often be referred to, e.g., for consistency checks of practical results, or for the assessment of different practical approaches for quantization. In addition to that, even though it does not exactly specify how to design a quantizer that has optimum performance, helpful guidelines can be learned from RDT that will be of high benefit in the following chapters.

2.1 Definition of the Rate Distortion Function

The RDT was developed principally to extend information theory to sources with continuous amplitudes. Numerous publications exist, e.g., [CT91], [Ber71], and [Yeu02]. The problem addressed by the RDT is illustrated in Figure 2.1.
Given is a set of N continuous independent and identically distributed (i.i.d.) random variables (random process)

$$\mathbf{X} := \begin{bmatrix} X_0 & X_1 & \cdots X_{N-1} \end{bmatrix}^T \tag{2.1}$$

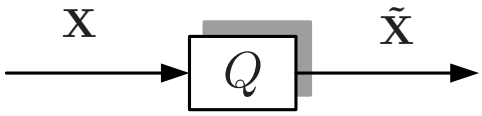


Figure 2.1: Quantization of sets of continuous random variables, $\mathbf{X}$.

(with variances $\sigma_0^2 = \sigma_1^2 = \cdots = \sigma_{N-1}^2 = \sigma^2$ and zero mean). The quantizer Q outputs a set of output random variables

$$\tilde{\mathbf{X}} := \begin{bmatrix} \tilde{X}_0 & \tilde{X}_1 & \cdots \tilde{X}_{N-1} \end{bmatrix}^T. \tag{2.2}$$

The output variables are a finite representation of the input random variables and therefore have discrete amplitudes. The quantizer Q is defined as the mapping of the (infinite) input alphabet $\mathcal{X}$ related to the random input variables $\mathbf{X}$ to the finite output alphabet $\tilde{\mathcal{X}}$ related to the output variables $\tilde{\mathbf{X}}$,

$$Q : \mathcal{X} \to \tilde{\mathcal{X}}. \tag{2.3}$$

The input random variables are characterized by the (multivariate) probability density function (PDF)

$$p(\mathbf{x}) := p_{\mathbf{X}}(\mathbf{x}). \tag{2.4}$$

In analogy to this, the set (realization of the random process) of discrete valued output variables is related to the (multivariate) probability mass function (PMF)

$$q(\tilde{\mathbf{x}}) := p_{\tilde{\mathbf{X}}}(\tilde{\mathbf{x}}). \tag{2.5}$$

2.1.1 Definition of a Quantization Cost Function

For an error-free representation of the continuous input random variables by the discrete output variables, an infinite number of representations and hence, if expressed in binary digits (bits), infinite bit rate would be necessary. Given only a finite number of representations, distortion d is introduced to rate the cost for representing a **single event** of a set of random variables, given as symbol $\mathbf{x} = \begin{bmatrix} x_0 & x_1 & \cdots & x_{N-1} \end{bmatrix}^T \in \mathcal{X}$, by a representative $\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}_0 & \tilde{x}_1 & \cdots & \tilde{x}_{N-1} \end{bmatrix}^T \in \tilde{\mathcal{X}}$,

$$d(\mathbf{x}, \tilde{\mathbf{x}}) = \frac{1}{N} \sum_{i=0}^{N-1} d(x_i, \tilde{x}_i). \tag{2.6}$$

The quantization cost for a single event of a set of N random variables is hence the average of the cost for each vector component based on the scalar distortion measure $d(x, \tilde{x})$. Several distortion measures have been investigated in the literature, for example the *Hamming criterion* [CT91], the *magnitude error criterion* [Ber71], and the *squared error criterion* [Ber71]. In most cases and also in the remainder of this thesis, rate distortion results are based on the squared error criterion

$$d(x_i, \tilde{x}_i) = (x_i - \tilde{x}_i)^2. \tag{2.7}$$

Considering the mapping of input to output variables, it is assumed that, given one event related to the set of the input random variables, the quantizer Q outputs that representation which minimizes the cost function (2.6).
Each quantizer Q has a fixed mapping of input variables to output variables and is characterized by the expectation of the quantization cost function,

$$D := E\{d(\mathbf{x}, \tilde{\mathbf{x}})\} = \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \int_{\mathbf{x} \in \mathcal{X}} p_{\mathbf{X}, \tilde{\mathbf{X}}}(\mathbf{x}, \tilde{\mathbf{x}}) \cdot d(\mathbf{x}, \tilde{\mathbf{x}}) \mathrm{d}\mathbf{x}. \tag{2.8}$$

This expectation value is a function of the joint PDF for input and output variables, which can be expressed in terms of the input PDF $p(\mathbf{x})$ and the conditional PDF $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$ which depends on quantizer Q,

$$p(\mathbf{x}, \tilde{\mathbf{x}}) := p_{\mathbf{X}, \tilde{\mathbf{X}}}(\mathbf{x}, \tilde{\mathbf{x}}) = p(\mathbf{x}) \cdot p_Q(\tilde{\mathbf{x}}|\mathbf{x}) \tag{2.9}$$

In the case of the squared error criterion for the distortion (2.7), the expectation of the cost function is denoted as the mean squared error (MSE).

2.1.2 Definition of the Information Rate

The information output rate of a quantizer is the *mutual information* (MI), defined, e.g., in [Ber71] as follows:

$$I(\mathbf{X}; \tilde{\mathbf{X}}) = h(\mathbf{X}) - h(\mathbf{X}|\tilde{\mathbf{X}}). \tag{2.10}$$

This mutual information is a function of the *differential entropy* related to the PDF of the set of input random variables with continuous amplitudes and the *conditional differential entropy* related to the conditional PDF $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$ which characterizes the quantizer. The *differential entropy* for a given amplitude continuous random variable and the corresponding PDF $p(x)$ is defined as

$$h(\mathbf{X}) = -\int_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \cdot \log_2(p(\mathbf{x})) \mathrm{d}\mathbf{x}. \tag{2.11}$$

Differential entropies are used here because both, $p(x)$ and $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$ are related to random variables with continuous amplitudes. The logarithm in (2.11) is to the base of two so that all information rates are specified as *bit rates* rather than in *nats* which is also very common in the literature.
The mutual information can be expressed as the *Kullback-Leibler* distance as [CT91]

$$I(\mathbf{X}; \tilde{\mathbf{X}}) = \sum_{\tilde{\mathbf{x}} \in \tilde{\mathcal{X}}} \int_{\mathbf{x} \in \mathcal{X}} p(\mathbf{x}) \cdot p_Q(\tilde{\mathbf{x}}|\mathbf{x}) \cdot \log_2 \frac{p_Q(\tilde{\mathbf{x}}|\mathbf{x})}{q(\tilde{\mathbf{x}})} \mathrm{d}\mathbf{x}. \tag{2.12}$$

In this equation, the mutual information is a function of the source distribution $p(\mathbf{x})$, the mapping function or conditional PDF $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$ which characterizes the quantizer, and the probability mass function $q(\tilde{\mathbf{x}})$ from (2.5). In the context of quantization, the mutual information is the expectation value of the number of bits (the bit rate) required to represent vectors $\mathbf{x}$ by the quantized representations $\tilde{\mathbf{x}}$ given a fixed source distribution ($p(\mathbf{x})$) and quantizer ($p_Q(\tilde{\mathbf{x}}|\mathbf{x})$).

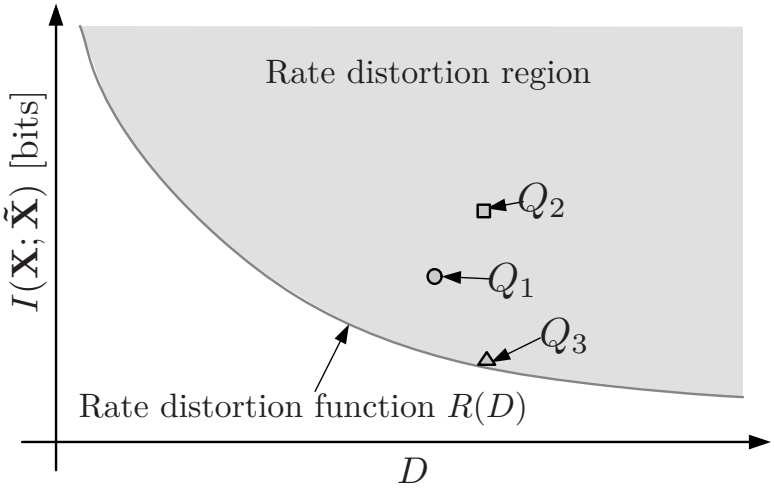


Figure 2.2: Rate distortion plane with quantizers Q_1, Q_2, and Q_3.

2.1.3 The *Rate Distortion Function* (RDF)

So far, nothing has been said about the quantization cost which is also a function of the quantizer and hence the conditional PDF $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$. An efficient quantizer is one that operates with a low bit rate and retains the expectation of the quantizer cost function as low as possible. These constraints, however, are contradictory according to (2.8) and (2.12) which shall be illustrated by the following observations:

- The quantization performance can be increased by allowing a larger number of representatives in the output alphabet. In this case, the expectation of the cost function (2.8) decreases. At the same time, with a higher number of representatives, naturally the average bit rate is increased.

- The average bit rate of the quantizer can be decreased by reducing the number of representatives. At the same time, with a lower number of representatives, the expectation of the quantizer cost function is of course increased.

In the design of the most efficient quantizer, the mapping function $p_Q(\tilde{\mathbf{x}}|\mathbf{x})$ must hence be chosen such that the best **cost-bit-rate-balance** is achieved. This is demonstrated by Figure 2.2. A plane spanned by the quantization cost function D on the x- and the MI (and hence bit rate) on the y-axis is shown qualitatively. For a given distribution of the input random variables, different quantizers are represented by different points of operation, characterized by a pair of MI and D. In the Figure, example quantizers are shown as Q_1, Q_2 and Q_3. All these quantizers are located in the *rate distortion region*, painted in gray color. Given a fixed allowed maximum distortion D, among all quantizers, there will be one with a minimum information rate. The corresponding point of operation is located on the edge of the rate distortion region and defines one point of the *rate distortion function* $R(D)$.
In a mathematical sense, the point of operation at which this quantizer is located

is defined as the minimum achievable mutual information (necessary average bit rate), given the side constraint that a maximum distortion D is not exceeded,

$$R(D) = \min_{p(\mathbf{x}|\tilde{\mathbf{x}}):\sum\int p(\mathbf{x})\cdot p_{\mathcal{Q}}(\tilde{\mathbf{x}}|\mathbf{x})\cdot d(\mathbf{x},\tilde{\mathbf{x}})\mathrm{d}\mathbf{x}\leq D} I(\mathbf{X};\tilde{\mathbf{X}}). \tag{2.13}$$

$R(D)$ is nonnegative, monotonic decreasing and convex [Ber71].

2.1.4 The *Distortion Rate Function* (DRF)

The definition of the rate distortion function for a given input random variable was based on the minimum information rate for a fixed distortion D. An optimization the other way around is also possible: Given a specific information rate R, among all quantizers the cost function $D(R)$ is minimal for one specific quantizer. The corresponding pair of rate and minimal distortion defines one point of the distortion rate function. Due to the properties of the rate distortion function, the distortion rate function can be calculated by inverting $R(D)$.

2.2 Calculation of the Rate Distortion Function

The rate distortion function can only be calculated explicitly in very specific cases. In particular, in case of the squared error distortion criterion (2.7), it is possible to calculate the rate distortion function for an i.i.d. Gaussian random variable X with zero-mean, variance σ^2, and with PDF

$$p_X(x) = \mathcal{N}(x, 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\frac{-x^2}{2\sigma^2}. \tag{2.14}$$

The rate distortion function is

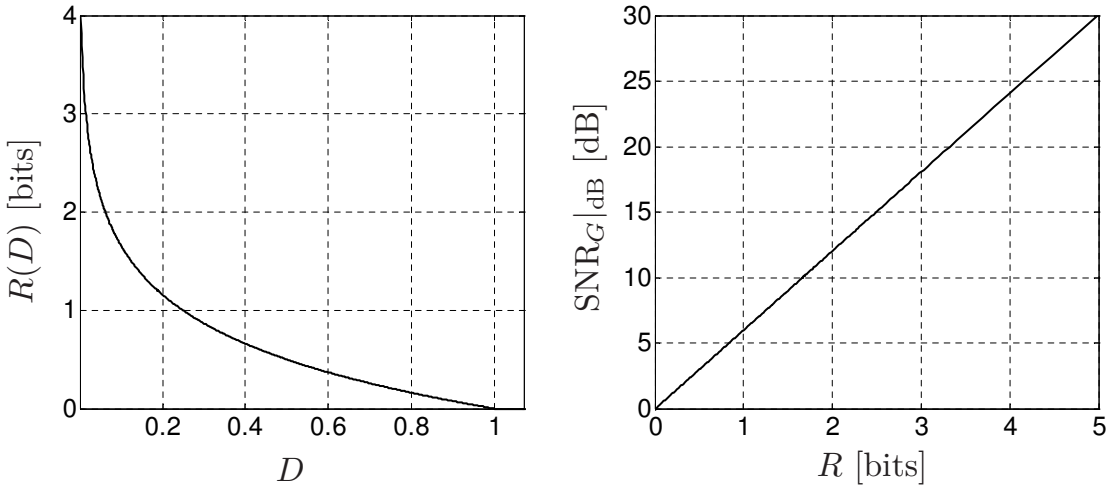
$$R_G(D) = \begin{cases} \frac{1}{2}\log_2\frac{\sigma^2}{D} & , 0 \leq D < \sigma^2 \\ 0 & , D \geq \sigma^2 \end{cases} \tag{2.15}$$

and is shown in Figure 2.3 a) for $\sigma^2 = 1$.
The distortion rate function is determined by inverting equation (2.15). Instead of calculating the distortion $D(R)$ (MSE), it is often useful to consider the normalized distortion $\mathcal{D}(R) = D(R)/\sigma^2$, the normalized logarithmic distortion in dB,

$$\mathcal{D}(R)|_{\mathrm{dB}} = 10 \cdot \log_{10}(\mathcal{D}(R)) = 10 \cdot \log_{10}(D(R)/\sigma^2), \tag{2.16}$$

or the logarithmic signal-to-noise ratio (SNR) in dB,

$$\mathrm{SNR}(R)|_{\mathrm{dB}} = 10\cdot\log_{10}(E\{X^2\}/D(R)) = 10\cdot\log_{10}(\sigma^2/D(R)) = -\mathcal{D}(R)|_{\mathrm{dB}}. \tag{2.17}$$



a) Rate distortion function, $\sigma^2 = 1$

b) Logarithmic SNR in dB over rate

Figure 2.3: Rate distortion function and respective SNR over bit rate for an i.i.d. Gaussian random variable. The logarithmic SNR in b) is calculated from the distortion rate function according to (2.17).

The specification of the logarithmic SNR is a common practice in quantization and will often be used in the remainder of this thesis.
For the Gaussian random variable, the logarithmic SNR is

$$\mathrm{SNR}_G(R)|_{\mathrm{dB}} = 10 \cdot \log_{10}(2^{2 \cdot R}) \approx 6.02 \cdot R. \tag{2.18}$$

This result is often referred to as the *6-dB-per-bit rule* and states that for each additional bit spent for quantization, the logarithmic SNR is increased by approximately 6 dB. The logarithmic SNR in dB according to the rate distortion function for a Gaussian input variable is shown in Figure 2.3 b). We will refer to the 6-dB-per-bit rule as the asymptotic quantization performance in various situations in this thesis.

2.2.1 Rate Distortion Bounds

For random variables with other distributions than the Gaussian, the rate distortion function can not be explicitly calculated. The *Shannon Lower Bound* (SLB) $R_{\mathrm{SLB}}(D)$, however, lowerbounds the rate distortion function $R(D)$ for arbitrary distributions. The derivation of this function is based on the fact that, given a fixed signal variance, the maximum differential entropy is generally attained for a Gaussian PDF [Ber71]. Taking advantage of this fact, a lower bound for the mutual information (2.10) can be calculated for a given PDF $p(\mathbf{x})$ and the respective differential entropy $h(\mathbf{X})$ as

$$R(D) = \min_{p(\mathbf{x}|\tilde{\mathbf{x}}):\sum \int p(\mathbf{x}) \cdot p_{\mathcal{Q}}(\tilde{\mathbf{x}}|\mathbf{x}) \cdot d(\mathbf{x},\tilde{\mathbf{x}}) \mathrm{d}\mathbf{x} \leq D} I(\mathbf{X};\tilde{\mathbf{X}}) \tag{2.19}$$

$$\geq h(\mathbf{X}) - \max_{p(\mathbf{x}|\tilde{\mathbf{x}}):\sum \int p(\mathbf{x}) \cdot p_{\mathcal{Q}}(\tilde{\mathbf{x}}|\mathbf{x}) \cdot d(\mathbf{x},\tilde{\mathbf{x}}) \mathrm{d}\mathbf{x} \leq D} h(\mathbf{X}|\tilde{\mathbf{X}}). \tag{2.20}$$

For the computation of the *Shannon Lower Bound*, the second part of this equation is upper bounded by the conditional differential entropy $h(D)$ of a random variable with variance D which is assumed to be Gaussian distributed, hence

$$R_{\mathrm{SLB}}(D) = h(\mathbf{X}) - h(D) \leq R(D). \tag{2.21}$$

In analogy to this, also an upper bound for the rate distortion function is given implicitly by (2.15),

$$R(D) \quad = \min_{p(\mathbf{x}|\tilde{\mathbf{x}}):\sum \int p(\mathbf{x})\cdot p_{\mathcal{Q}}(\tilde{\mathbf{x}}|\mathbf{x})\cdot d(\mathbf{x},\tilde{\mathbf{x}})\mathrm{d}\mathbf{x}\leq D} I(\mathbf{X};\tilde{\mathbf{X}}) \leq R_G(D). \tag{2.22}$$

The rate distortion function $R(D)$ for a given PDF and respective differential entropy $h(\mathbf{X})$ is hence bounded by (2.21) and (2.22).

$$R_{\mathrm{SLB}}(D) \leq R(D) \leq R_G(D). \tag{2.23}$$

The Shannon Lower Bound can also be transformed into the distortion rate function by inversion of (2.21),

$$D_{\mathrm{SLB}}(R)|_{h(\mathbf{X})} = 2^{2\cdot(h(\mathbf{X})-R)}. \tag{2.24}$$

The distortion rate function for a Gaussian variable in relation to the SLB for a random variable with arbitrary distribution is defined as

$$\Delta_{D,\mathbf{X}}|_{\mathrm{dB}} = 10\cdot\log_{10}\left(\frac{D_G(R)}{D_{\mathrm{SLB}}(R)|_{h(\mathbf{X})}}\right) = 10\cdot\log_{10}\left(\frac{\frac{1}{2}\cdot\log_2(2\pi e\sigma^2)}{h(\mathbf{X})}\right) \tag{2.25}$$

and given for different PDFs in the literature, e.g., [Erd04].

2.2.2 Approximation by *Blahuts Method*

The Shannon Lower Bound defines only the asymptotic behavior of the rate distortion function for large values of R. A way to calculate the rate distortion function also for low information rates was proposed in [Bla72], [Ari72], and [Csi74] and is known as *Blahuts Method*. Results for the approximation of the rate distortion function based on Blahuts Method were shown in [NZ78] also for signals with non-Gaussian distribution.
According to Blahuts Method, the solution of the problem in (2.13) can be found by constructing isolated function values of the rate distortion curve following an iterative approach. The function values are found for different values of the parameter s which is the gradient of the rate distortion curve. A unique result is guaranteed because the rate distortion function is convex.

2.3 Rate Distortion Function for Stationary Correlated Gaussian Sources

For the calculation of the rate distortion function for correlated sources (intraframe correlation), a set of N independent continuous random variables with different variances

$$\sigma_i^2 \neq \sigma_j^2 \qquad \forall \qquad i \neq j, \qquad i,j \in \{0,\ldots,N-1\} \tag{2.26}$$

is considered at first which produces events of vectors $\mathbf{x}$ being subject to quantization. In order to explicitly **calculate** the rate distortion function, it is assumed that the random variables have Gaussian distribution.
For each of the N random variables X_i, the individual information rate is R_i and the quantizer cost function D_i. Since in quantization, often an effective bit rate per sample rather than the bit rate per vector is specified (e.g. in (3.37)), the **average information rate and distortion per vector coordinate** shall be computed as the mean over all random variables in the following,

$$\bar{R} = \frac{1}{N} \cdot \sum_{i=0}^{N-1} R_i \tag{2.27}$$

$$\bar{D} = \frac{1}{N} \sum_{i=0}^{N-1} D_i. \tag{2.28}$$

According to the optimization procedure described in, e.g., [CT91], a new variable D_0 is introduced and the distortion for each random variable is

$$D_i = \begin{cases} D_0 & \text{, if } D_0 < \sigma_i^2 \\ \sigma_i^2 & \text{, if } D_0 \geq \sigma_i^2 \end{cases}. \tag{2.29}$$

The corresponding information rate can be calculated in analogy to 2.15 as

$$R_i = \begin{cases} \frac{1}{2} \cdot \log_2(\frac{\sigma_i^2}{D_0}) & \text{, if } D_0 < \sigma_i^2 \\ 0 & \text{, if } D_0 \geq \sigma_i^2 \end{cases}. \tag{2.30}$$

Both equations define that the individual bit rates are setup such that the same quantization distortion is introduced for all random variables except for those for which the introduced quantization distortion would be greater than the variance of the random variable. In the latter case, no bits are reserved for the quantization of the individual random variable. The described procedure is called *reverse waterfilling* [CT91]. Similar to Blahuts Method, the rate distortion curve can be constructed pointwise based on pairs of R and D which are computed for the slope

s of the rate distortion curve. According to [Ber71], D_0 is computed for a given value of s as

$$D_0 = -\frac{1}{2s}. \tag{2.31}$$

Consequently, the average distortion is

$$\bar{D} = \frac{1}{N}\sum_{i=0}^{N-1} \min(D_0, \sigma_i^2), \tag{2.32}$$

and the corresponding information rate in bits is

$$\bar{R}(\bar{D}) = \frac{1}{N}\sum_{i=0}^{N-1} \max(0, \frac{1}{2}\log_2(\frac{\sigma_i^2}{D_0})). \tag{2.33}$$

$\bar{R}(\bar{D})$ is a function of $\bar{D}$ since D_0 in (2.33) can be expressed as a function of $\bar{D}$ on the basis of (2.32).

2.3.1 Asymptotic Behavior for High Bit Rates

For high bit rates, it can be assumed that $\min(D_0, \sigma_i^2) = D_0$ for all i. In this case, the rate distortion function is

$$\bar{R}(\bar{D}) = \frac{1}{2N} \cdot \sum_{i=0}^{N-1} \log_2(\frac{\sigma_i^2}{D_0}), \tag{2.34}$$

and the distortion is

$$\bar{D} = D_0. \tag{2.35}$$

With (2.35) and the inversion of equation (2.34) to compute $\bar{D}(\bar{R})$ as well as the mean of the variances of all independent random variables,

$$\overline{\sigma^2} = \frac{1}{N}\sum_{i=0}^{(N-1)} \sigma_i^2, \tag{2.36}$$

the asymptotic SNR for high bit rates in dB is derived as

$$\begin{aligned} \mathrm{SNR}(\bar{R})|_{\mathrm{dB}} &= 10 \cdot \log_{10}\left(\frac{\overline{\sigma^2}}{\bar{D}(\bar{R})}\right) \\ &= 10 \cdot \log_{10}\left(\frac{\frac{1}{N}\sum_{i=0}^{(N-1)} \sigma_i^2}{(\prod_{i=0}^{N-1} \sigma_i^2)^{\frac{1}{N}} \cdot 2^{-2\bar{R}}}\right) \\ &= 10 \cdot \log_{10}\left(\frac{\frac{1}{N}\sum_{i=0}^{N-1} \sigma_i^2}{\exp(\frac{1}{N}\sum_{i=0}^{N-1} \ln(\sigma_i^2))}\right) + 6.02 \cdot \bar{R}. \end{aligned} \tag{2.37}$$

2.3.2 Decorrelation by Singular Value Decomposition (SVD)

The case of N **independent** random variables with different variances can be adapted for the case of a set of N **dependent** random variables if the following constraints are fulfilled:

- The cross correlation function (CCF) of two random variables with index i and j from the overall set is

$$\varphi_{X_i,X_j} = E\{X_i \cdot X_j\}. \tag{2.38}$$

- Considering two pairs of random variables from the overall set with index i_0 and $j_0 = i_0 + \Delta_i$ for the one and i_1 and $j_1 = i_1 + \Delta_i$ for the other pair with $i_0 \neq i_1$, the cross correlation function is only a function of $|\Delta_i|$:

$$\varphi_{X,X}(|\Delta_i|) := \varphi_{X_{i_0},X_{j_0}} = \varphi_{X_{i_1},X_{j_1}} \ \forall\ i_0, i_1, \Delta_i. \tag{2.39}$$

These constraints are fulfilled for, e.g., ergodic sequences X_t related to a stationary correlated Gaussian variable with zero mean, recorded at equidistant time intervals $t = T_0, T_1, \cdots$ so that the set of random variables is defined as $\mathbf{X} = \begin{bmatrix} X_{t=T_0} & X_{t=T_1} & \cdots & X_{t=T_{N-1}} \end{bmatrix}^T$ [Ber71]. The corresponding multivariate PDF is

$$p_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \mathbf{\Phi_X}) = \frac{1}{(2\pi)^{N/2}} \cdot \frac{1}{\sqrt{\det(\mathbf{\Phi_X})}} \cdot \exp(-\frac{1}{2}\mathbf{x} \cdot \mathbf{\Phi}_{\mathbf{X}}^{-1} \cdot \mathbf{x}^T) \tag{2.40}$$

with the *covariance matrix*

$$\mathbf{\Phi_X} = \begin{bmatrix} \varphi_{X,X}(0) & \varphi_{X,X}(1) & \cdots & \varphi_{X,X}(N-1) \\ \varphi_{X,X}(1) & \varphi_{X,X}(0) & \cdots & \varphi_{X,X}(N-2) \\ \cdots & \cdots & \cdots & \cdots \\ \varphi_{X,X}(N-1) & \varphi_{X,X}(N-2) & \cdots & \varphi_{X,X}(0) \end{bmatrix} \tag{2.41}$$

composed of the *autocorrelation function* (ACF) values $\varphi_{X,X}(i)$. The term $\det(\mathbf{\Phi_X})$ refers to the determinant of the covariance matrix $\mathbf{\Phi_X}$.
In [Ber71] it is shown that $\mathbf{\Phi_X}$ is a *symmetric Toeplitz matrix* if the constraints specified earlier are fulfilled [GS58]. Another matrix

$$\mathbf{\Gamma_X} = \begin{bmatrix} \mathbf{x}_{e_0} & \mathbf{x}_{e_1} & \cdots \mathbf{x}_{e_{N-1}} \end{bmatrix} \tag{2.42}$$

can be calculated for a given matrix $\mathbf{\Phi_X}$ that is composed of the N orthogonal Eigenvectors $\mathbf{x}_{e_i}$. Based on $\mathbf{\Gamma_X}$, $\mathbf{\Phi_X}$ can be decomposed by means of a Singular Value Decomposition (SVD) as proposed in the context of the *Karhunen Loeve Transform* (KLT)[Kar47] according to

$$\mathbf{\Phi_X} = \mathbf{\Gamma_X} \cdot \mathbf{\Lambda_X} \cdot \mathbf{\Gamma}_{\mathbf{X}}^T, \tag{2.43}$$

to produce the diagonal matrix $\mathbf{\Lambda_X}$ which is composed of the Eigenvalues $\lambda_{\mathbf{X},i}$,

$$\mathbf{\Lambda_X} = \begin{bmatrix} \lambda_{\mathbf{X},0} & 0 & \cdots & 0 \\ 0 & \lambda_{\mathbf{X},1} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_{\mathbf{X},N-1} \end{bmatrix}. \tag{2.44}$$

Due to the diagonalization of matrix $\mathbf{\Phi_X}$, the problem related to a set of dependent Gaussian random variables has been transformed into a problem related to a set of N independent random variables with different variances in the coordinate system of the *principal axes* (Eigenvectors $\mathbf{x}_{e_i}$). The Eigenvalues $\lambda_{\mathbf{X},i}$ are the equivalent of the different squared variances of the independent random variables. In analogy to (2.37), the asymptotic SNR for high bit rates for a set of dependent Gaussian random variables is hence

$$\mathrm{SNR}(\bar{R})|_{\mathrm{dB}} = 10 \cdot \log_{10}\left(\frac{\frac{1}{N}\sum_{i=0}^{N-1}\lambda_{\mathbf{X},i}}{\exp\left(\frac{1}{N}\sum_{i=0}^{N-1}\ln(\lambda_{\mathbf{X},i})\right)}\right) + 6.02 \cdot \bar{R} \tag{2.45}$$

An example that well illustrates the diagonalization of an example covariance matrix is shown in Section 3.2.1.3 for the PDF related to a two-dimensional correlated Gaussian random variable.

2.3.3 Toeplitz Distribution Theorem

Given an ergodic sequence originating from a stationary Gaussian random variable X, it was shown in [GS58] that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=0}^{N-1}\mathcal{G}(\lambda_{\mathbf{X},k}) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathcal{G}(\phi_{\mathbf{X}}(\Omega))\mathrm{d}\Omega \tag{2.46}$$

with $\mathcal{G}$ being an arbitrary continuous function, the Eigenvalues $\lambda_{\mathbf{X},i}$ of the covariance matrix $\mathbf{\Phi_X}$, and the power spectral density (PSD), e.g., [VM06],

$$\phi_{\mathbf{X}}(\Omega) = \sum_{k=-\infty}^{\infty}\varphi_{X,X}(k)\cdot\exp(-jk\Omega) \tag{2.47}$$

computed from the autocorrelation function (ACF) for variable X_t. With respect to the reverse waterfilling (2.32 and 2.33), the average distortion and rate can be computed from the PSD as

$$\bar{D} = \frac{1}{2\pi}\cdot\int_{-\pi}^{\pi}\min(D_0, \phi_{\mathbf{X}}(\Omega))\mathrm{d}\Omega \tag{2.48}$$

and

$$\bar{R}(\bar{D}) = \frac{1}{4\pi} \cdot \int_{-\pi}^{\pi} \max(0, \log_2(\frac{\phi_{\mathbf{X}}(\Omega)}{D_0}) \mathrm{d}\Omega, \tag{2.49}$$

respectively. The SNR related to the asymptotic rate distortion function for a correlated Gaussian source can be hence calculated from the PSD as

$$\mathrm{SNR}(\bar{R})|_{\mathrm{dB}} = 10 \cdot \log_{10}(\frac{\frac{1}{2\pi} \cdot \int_{-\pi}^{\pi} \phi_{\mathbf{X}}(\Omega) \mathrm{d}\Omega}{\exp(\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln(\phi_{\mathbf{X}}(\Omega)) \mathrm{d}\Omega}) + 6.02 \cdot \bar{R}. \tag{2.50}$$

for high bit rates and as the limit for $N \to \infty$ (2.46). In anticipation of Section 5.2.4.2, the first part of (2.50) is equal to the spectral flatness measure (SFM) given in (5.67).

2.3.4 Example SNR Plot for Correlated Gaussian Sources

In Figure 2.4, example plots for the SNR according to the rate distortion function for correlated Gaussian sources are depicted for three different covariance matrices and hence Eigenvalues as the solid line with circle markers, the dashed line with square markers, and the dotted line with triangle markers. The matrices $\mathbf{\Lambda}_{\mathbf{X},a}$, $\mathbf{\Lambda}_{\mathbf{X},b}$, and $\mathbf{\Lambda}_{\mathbf{X},c}$ for the examples have been constructed such that the asymptotic SNR for high bit rates is the same in all cases while the behavior for lower bit rates is significantly different due to different reverse waterfilling characteristics. Pairs of rate and distortion have been computed for a number of values for parameter s according to Section 2.3 with the Eigenvalues as the variances in (2.32) and (2.33). In addition to the mentioned curves, the SNR related to the rate distortion function for an uncorrelated Gaussian i.i.d. random variable ($\mathbf{\Lambda}_{\mathbf{X}} = \mathbf{E}$) is shown. The asymptotic difference between RDT for correlated and uncorrelated Gaussian sources for high bit rates is

$$\Delta_{\mathrm{SNR}}|_{\mathrm{dB}} = 10 \cdot \log_{10}(\frac{\frac{1}{N} \sum_{i=0}^{N-1} \lambda_{\mathbf{X},i}}{\exp(\frac{1}{N} \sum_{i=0}^{N-1} \ln(\lambda_{\mathbf{X},i}))}) = 10 \text{ dB} \tag{2.51}$$

for $\mathbf{\Lambda}_{\mathbf{X},a}$, $\mathbf{\Lambda}_{\mathbf{X},b}$, and $\mathbf{\Lambda}_{\mathbf{X},c}$.
It will be shown in later chapters of the thesis that Δ_{SNR} is identical to the (asymptotic) memory advantage in vector quantization and the (asymptotic) maximum prediction gain in linear predictive coding which will be introduced in Section 3 and 5, respectively.
Bit rates at which the reverse waterfilling has an impact on the performance are those where the SNR plots deviate from the (linear) continuation of the asymptotic RDT curve for high bit rates (gray line with label $\mathrm{SNR}_{\mathrm{hr}}$), e.g., $\bar{R}_{\mathrm{low},\mathbf{\Lambda}_{\mathbf{X},a}}$ in Figure

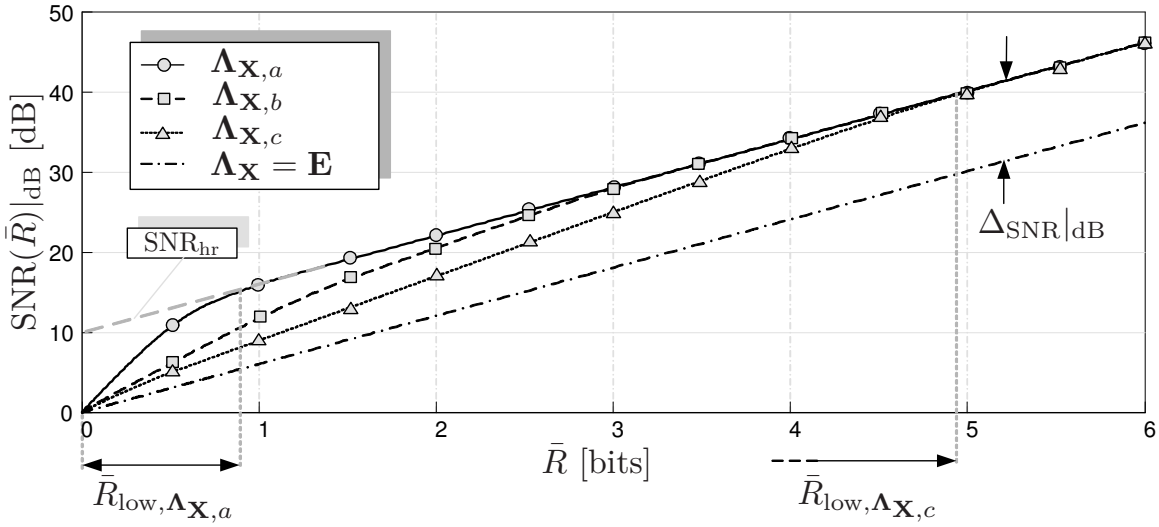


Figure 2.4: SNR curves related to the rate distortion function for correlated Gaussian sources (different covariance and hence Eigenvalue matrices $\mathbf{\Lambda}_{\mathbf{X},i}$ as listed in Appendix B.2 to demonstrate different reverse waterfilling behaviors).

2.4. Accordingly, a new definition of "low bit rate areas" and "high bit rate areas" can be derived which we will refer to in Section 5.2.5.3. From the plots it is obvious that this new definition of high and low bit rate areas depends strongly on the signal, e.g.,

$$R_{\mathrm{low},\mathbf{\Lambda}_{\mathbf{X},a}} \approx 0.8 \text{ bits} \qquad \text{and} \qquad R_{\mathrm{low},\mathbf{\Lambda}_{\mathbf{X},c}} \approx 5 \text{ bits}. \tag{2.52}$$

Especially the low bit rate areas will be subject of additional theoretical investigations in Section 5.2.6.1 and 5.2.6.2. As a conclusion of these investigations, a new solution to realize the reverse waterfilling in combined linear prediction and quantization will be derived in Section 5.
The Eigenvalue matrices for the three example curves, $\mathbf{\Lambda}_{\mathbf{X},a}$, $\mathbf{\Lambda}_{\mathbf{X},b}$, and $\mathbf{\Lambda}_{\mathbf{X},c}$ are listed in detail in Appendix B.2.

3

Quantization

In the previous chapter, results from the rate distortion theory were reviewed. Bounds on the maximum theoretically achievable quantization performance were discussed which will be helpful in the assessment of approaches for quantization in practice. Nevertheless, rate distortion theory refuses to present practical recipes how to approach the given bounds, not least because all bounds are based on the (unrealistic) assumption of infinite block lengths.

In this chapter, quantization will be investigated from a practical point of view, in the literature denoted as *asymptotic quantization theory*. Different quantization schemes will be investigated, and the corresponding *operational rate distortion functions* will be put in relation to the results from rate distortion theory. Various text books and other publications have been written on quantization, e.g., [Neu96], [JN84], [Abu90], and [GR92]. A very good overview on all aspects of quantization is given in [GN98].

This thesis addresses the problem of quantization of audio signals (audio coding). In most practical applications, audio coding is done purely in the digital domain since the analog input signal has already been transformed into a high precision digital representation with a large number of bits in the analog-to-digital converter (ADC). Quantization in this context should be considered a *requantization* operation to find a representation of the digital signal with a lower number of bits rather than an analog-to-digital conversion. The representation produced by the ADC nowadays, however, is of such high precision in practice (e.g. 24 or 32 bits) that the digital signals can be treated as signals with continuous amplitudes. Nevertheless, due to the preceding ADC, the signal amplitudes are guaranteed to be of limited amplitude so that quantization overload effects, if happening, can not be influenced by the digital audio codec in most cases and will not be considered here.

In the first part of this chapter, the principles of scalar quantization (SQ) will be briefly reviewed. It will be distinguished between *fixed and variable rate coding*. The performance achievable by the proposed scalar quantizers will be presented for uncorrelated (also referred to as *memoryless*) signals only, approaches to exploit the correlation immanent to a signal with memory will be subject of Chapter 5.

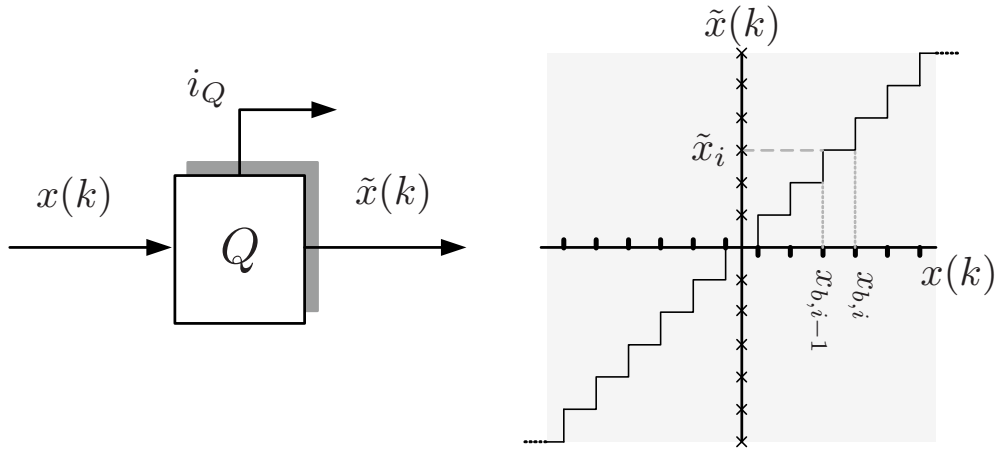


Figure 3.1: Scalar (uniform) quantization of amplitude continuous signal $x(k)$.

In the second part of this chapter, the results for SQ will be generalized for vector quantization (VQ). It will be shown that VQ has advantages compared to SQ and asymptotically reaches the rate distortion function for infinite dimensions. The basis for all calculations of quantization performances in this chapter are statistical processes and random variables. In order to be applicable also for sequences of signal amplitudes (and hence audio signals), it is assumed that all signals to be quantized are stationary and ergodic.

3.1 Scalar Quantization (SQ)

In scalar quantization (SQ), at a given time index k, an input signal $x := x(k) \in \mathcal{X}$ is mapped to an output signal $\tilde{x} := \tilde{x}(k) \in \tilde{\mathcal{X}}$. The input alphabet $\mathcal{X}$ is a (quasi) infinite set of signal samples with (quasi) continuous amplitudes, whereas the output alphabet $\tilde{\mathcal{X}}$ is a finite set of signal samples with discrete amplitude. The principle is shown on the left side in Figure 3.1. In order to be precise, note that the SQ is part of the encoder and outputs an index i_Q which is transformed into the reconstructed value $\tilde{x} = \tilde{x}_{i_Q}$ taken from $\tilde{\mathcal{X}}$ in the decoder. For the sake of simplicity, however, quantizer, index mapping, and value reconstruction will be considered as one unique entity and can be described as

$$Q : x \mapsto \tilde{x}. \tag{3.1}$$

A scalar quantizer is characterized by its $(N_Q + 1)$ *quantizer interval bounds* $(x_{b,-1}, x_{b,0}, \ldots, x_{b,N_Q-1})$ and the N_Q *quantizer amplitude reconstruction levels* $\tilde{x}_i \in \tilde{\mathcal{X}}$ with $i = 0, \ldots, (N_Q - 1)$. Each *quantization interval* $C_{\mathrm{sq},i}$ is defined as the space between neighboring interval bounds,

$$C_{\mathrm{sq},i} := \{x \in \mathbb{R} : x_{b,i-1} \leq x < x_{b,i}\} \tag{3.2}$$

with the corresponding quantization reconstruction level $\tilde{x}_i$ to be commonly located inside the interval. An example curve to map the input signal to the output signal

for a uniform scalar quantizer is shown on the right side of Figure 3.1. Assuming an unbounded input signal $x(k)$, the outer most quantization cells are also unbounded, $x_{b,-1} = -\infty$ and $x_{b,(N_Q-1)} = \infty$. The unbounded cells are related to the *quantization overload*. All other cells comprise the so-called *support* region of the quantizer. Given the PDF $p(x) := p_x(x)$ related to input signal x, the expectation of the quantization cost function or distortion is

$$D(x, Q(x)) = E\{d(x, Q(x))\} = \sum_{i=0}^{N_Q-1} \int_{C_{\mathrm{sq},i}} d(x, \tilde{x}_i) \cdot p(x)\mathrm{d}x \tag{3.3}$$

$$= \sum_{i=0}^{N_Q-1} \int_{C_{\mathrm{sq},i}} (x - \tilde{x}_i)^2 \cdot p(x)\mathrm{d}x \tag{3.4}$$

with the squared error criterion known from (2.7).

3.1.1 Fixed Rate SQ

In fixed rate SQ, given N_Q quantization reconstruction levels, the bit rate of $R_{\mathrm{fr}} = \log_2(N_Q)$ bits is constant for all times k. In the history of quantization, the target in the design of fixed rate SQ was mainly to reduce the quantization distortion for a specific distribution of the input signal, PDF $p(x)$, and a fixed value R_{fr}.

3.1.1.1 Uniform SQ

In Uniform Quantization, the range of quasi continuous amplitudes of the input signal is subdivided into N_Q intervals of equal size

$$\Delta_{\mathrm{u}} = \frac{2x_{\max}}{N_Q} \tag{3.5}$$

For a sufficiently large number N_Q, the input signal is assumed to be uniformly distributed within each quantization interval. Also, the distribution of signal $x(k)$ is assumed to be symmetric and limited by $-x_{\max} \leq x \leq x_{\max}$ (overload effects are neglected). The quantization reconstruction levels are located in the center of gravity with respect to the signal PDF within each interval and therefore the interval mid-points,

$$\tilde{x}_i = \frac{x_{b,i} - x_{b,i-1}}{2} \quad \forall \quad i = 0, \ldots, (N_Q - 1). \tag{3.6}$$

The introduced quantization distortion is

$$D(x, Q(x)) \cong \frac{\Delta_{\mathrm{u}}^2}{12} \tag{3.7}$$

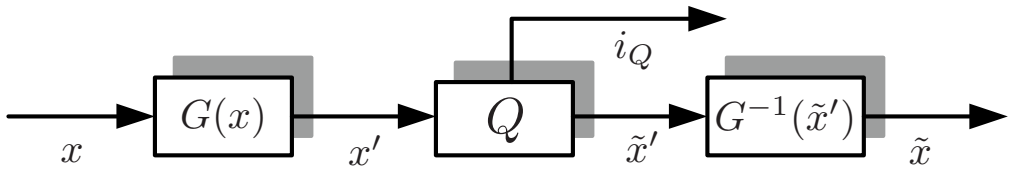


Figure 3.2: Non-Uniform SQ based on compression function $G(x)$ and uniform scalar quantizer Q.

[Ben48][VM06][JN84]. Correspondingly, the SNR is

$$\mathrm{SNR}_{\mathrm{sq,u}} = \frac{E\{x^2\}}{E\{(x - Q(x))^2\}} \cong 12 \cdot \frac{E\{x^2\}}{\Delta_{\mathrm{u}}^2} \tag{3.8}$$

Writing Δ_{u} as a function of the bit rate R_{fr}, the logarithmic SNR in decibel is

$$\mathrm{SNR}_{\mathrm{sq,u}}(R)|_{\mathrm{dB}} \cong \underbrace{10 \cdot \log_{10}(3 \cdot \frac{E\{x^2\}}{x_{\max}^2})}_{\Delta_{\mathrm{SNR,u}}|_{\mathrm{dB}}} + 6.02 \cdot R_{\mathrm{fr}}. \tag{3.9}$$

The second part of this equation is in accordance to the 6-db-per-bit rule from (2.18), and the first part is a constant offset $\Delta_{\mathrm{SNR,u}}$ which depends on the PDF of the input signal. Example PDF types and corresponding constants for uniform quantization are, e.g., given in [JN84].

3.1.1.2 Non-Uniform SQ

In non-uniform SQ, the quantization intervals $C_{\mathrm{sq},i}$ are no longer of equal size. One way to realize a non-uniform quantizer resolution is to use a non-linear compression function $G(x)$ in the first and to uniformly quantize the compressed output in the second step as shown in Figure 3.2. The reconstructed signal is $\tilde{x} = G^{-1}(Q(G(x)))$ with $G^{-1}(\tilde{x}')$ as the inverse of $G(x)$ so that $G^{-1}(G(x)) = x$. With the derivative of the compressor function, $G'(x) = \frac{\mathrm{d}G(x)}{\mathrm{d}x}$, it is shown in [Ben48] that for large values of N_Q, the quantization distortion is

$$D(x, Q(x)) \cong \frac{\Delta_{\mathrm{u}}^2}{12} \int_{-x_{\max}}^{x_{\max}} \frac{p(x)}{(G'(x))^2} \mathrm{d}x \tag{3.10}$$

(*Bennett's Integral*) with Δ_{u} related to the uniform quantizer (3.5). The derivative of the compressor function can also be interpreted in the context of a quantization reconstruction level density:
Given the constant *normalized point density* of the uniform quantizer Q with respect to signal x' as

$$\lambda_0(x') = \frac{1}{2 \cdot x'_{\max}} = \mathrm{const}, \tag{3.11}$$

due to the compressor function the point density with respect to signal x is

$$\lambda_1(x) = G'(x) \cdot \frac{1}{2 \cdot x'_{\max}}. \tag{3.12}$$

(3.10) can hence be written as

$$D(x, Q(x)) \approx \frac{1}{12} \frac{1}{N_Q^2} \int_{-x_{\max}}^{x_{\max}} \frac{p(x)}{(\lambda_1(x))^2} \mathrm{d}x \tag{3.13}$$

3.1.1.3 Optimal Non-Uniform SQ

For the assumption of high bit rates, given the PDF of the input signal as $p(x)$, the normalized point density function $\lambda(x)$ which is optimal in the sense to minimize the quantizer distortion [PD51] is

$$\lambda(x) = \frac{(p(x))^{1/3}}{\int_{-x_{\max}}^{x_{\max}} (p(y))^{1/3} \mathrm{d}y}. \tag{3.14}$$

Combining this result with (3.13) yields the *Panter and Dite Formula*

$$D(R_{\mathrm{fr}}) \cong \frac{1}{12} \left(\int_{-x_{\max}}^{x_{\max}} (p(x))^{1/3} \mathrm{d}x \right)^3 \cdot 2^{-2 \cdot R_{\mathrm{fr}}} \tag{3.15}$$

Writing this as the logarithmic SNR in decibel yields

$$\mathrm{SNR}_{\mathrm{sq,nu}}(R_{\mathrm{fr}})|_{\mathrm{dB}} \cong \underbrace{10 \cdot \log_{10} \left(12 \cdot \frac{E\{x^2\}}{\left(\int_{-x_{\max}}^{x_{\max}} p^{1/3}(x) \mathrm{d}x \right)^3} \right)}_{\Delta_{\mathrm{SNR,nu}}|_{\mathrm{dB}}} + 6.02 \cdot R_{\mathrm{fr}} \tag{3.16}$$

Again, besides the 6-db-per-bit rule a constant offset $\Delta_{\mathrm{SNR,nu}}$ is identified which is a function of the PDF of the input signal. Values for this offset for *Gamma*, *Laplacian*, *Gaussian*, and *uniform* distributions are, e.g., presented in [JN84]. The constant offset for a Gaussian distributed random variable is

$$\Delta_{\mathrm{SNR,nu,G}}|_{\mathrm{dB}} = 10 \cdot \log_{10}(\frac{12}{6\pi\sqrt{3}}) = -4.34\mathrm{dB}. \tag{3.17}$$

3.1.1.4 Lloyd-Max Quantization (LMQ)

So far, the results for non uniform SQ were based on the assumption of high bit rates. For the case of lower bit rates, an iterative optimization procedure to construct scalar non uniform quantizers has been proposed in [Llo82] and [Max60], nowadays referred to as the Lloyd-Max Quantizer (LMQ). In the iterative optimization procedure, quantizer threshold values and reconstruction levels are iteratively constructed to finally produce a quantizer which is locally or globally optimal. In [JN84], threshold values and reconstruction levels are given for bit rates of $R_{\text{fr}} = 1, 2, 3, 4$ bits for *Gamma*, *Laplacian*, *Gaussian*, and *uniform* distributed random variables. An important result related to both, the high rate PDF optimized SQ and the LMQ, is that

- the optimal positions of the quantizer reconstruction levels are in the center of gravity of the corresponding quantization intervals.
- the average quantization cost for each cell is roughly the same (also called the "Partial Distortion Theorem" [PD51]).

3.1.1.5 Logarithmic Non-Uniform SQ

According to (3.13), given the PDF $p(x)$ and a normalized point density $\lambda(x)$, the SNR is

$$\text{SNR}_{\text{sq,nu}} = 12 \cdot N_Q^2 \frac{\int_{-x_{\max}}^{x_{\max}} x^2 p(x) \mathrm{d}x}{\int_{-x_{\max}}^{x_{\max}} p(x)/(\lambda(x))^2 \mathrm{d}x} \tag{3.18}$$

For a choice of

$$\lambda(x) \sim \frac{1}{x}, \tag{3.19}$$

this SNR is independent from the PDF of the input signal. With respect to (3.12), this can be achieved by setting

$$G(x) = K_0 \cdot \ln(x) + K_1 \tag{3.20}$$

As a logarithmic compression curve is impractical for amplitudes close to zero, in *A-Law* SQ [Cat69], the compression curve is defined as

$$G(x) = \begin{cases} \frac{A|x|}{1+\ln(A)} \cdot \text{sign}(x) & \text{for } 0 \leq \frac{|x|}{x_{\max}} \leq \frac{1}{A} \\ x_{\max} \cdot \frac{1+\ln(A\frac{|x|}{x_{\max}})}{1+\ln(A)} \cdot \text{sign}(x) & \text{for } \frac{1}{A} < \frac{|x|}{x_{\max}} \leq 1. \end{cases} \tag{3.21}$$

In this context, the constant A defines a threshold value $x = \frac{x_{\max}}{A}$ at which the compression curve switches from linear for small amplitudes to logarithmic for high amplitudes. A very similar concept is known as μ-*Law* quantization [Hol49]

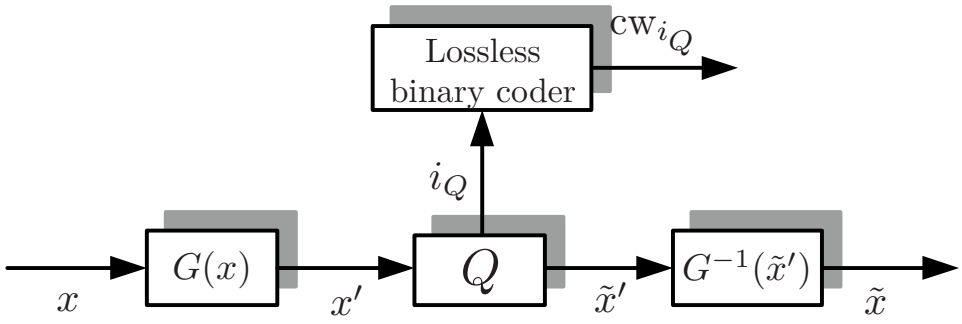


Figure 3.3: Scalar Quantization with lossless binary coder.

[PD51][Smi57]. The ITU-T G.711 speech codec [ITU88a] nowadays employed in ISDN [ITU88c] and VoIP [Bad06] is based on A-Law quantization with 8 bits and a value of $A = 87.56$ in Europe (μ-Law is used in Northern America and Japan). Due to the threshold functionality, the constant A is related to the dynamic range in which the quantizer produces a constant SNR, but at the same time also has an impact on the logarithmic SNR:

$$\text{SNR}_{\text{sq,nu,A}}|_{\text{dB}} = \underbrace{4.77 - 20 \cdot \log_{10}(1 + \ln(A))}_{\Delta_{\text{SNR,nu,A}}|_{\text{dB}}} + 6.02 \cdot R_{\text{fr}}. \tag{3.22}$$

Given a value of $A = 87.56$ as used in the G.711 codec, the dynamic range of the quantizer is 38 dB, e.g., [VM06], and the constant offset of the quantizer SNR in relation to the 6-dB-per-bit rule is

$$\Delta_{\text{SNR,nu,A}}|_{\text{dB}} = -9.99 \text{ dB}. \tag{3.23}$$

3.1.2 Variable Rate SQ

For variable bit rate SQ, Figure 3.2 is extended by the additional *lossless binary coder* block as shown in Figure 3.3. Compared to fixed rate SQ, in variable rate SQ, the index i_Q is fed into the lossless binary coder to be translated into a binary codeword cw_{i_Q} of variable instantaneous bit width $R_{\text{vr}}(k)$ prior to a transmission to the decoder.
It is known from information theory [Sha48], that a non-uniform distribution of the quantization indices i_Q can be exploited to achieve an average bit rate $\bar{R}_{\text{vr}}$ which is lower than that in fixed rate quantization (given the same number N_Q),

$$\bar{R}_{\text{vr}} = E\{R_{\text{vr}}(k)\} \leq R_{\text{fr}} = \log_2(N_Q) \tag{3.24}$$

Given the PMF $q(\tilde{x}_i)$ related to the quantization reconstruction level $\tilde{x}_i$, the minimum achievable average bit rate for variable rate quantization is given as the entropy of the quantizer output

$$\bar{R}_{\text{vr,min}} = H(Q(x)) = -\sum_{i=0}^{N_Q-1} q(\tilde{x}_i) \log_2(q(\tilde{x}_i)). \tag{3.25}$$

Variable rate quantization is often denoted as *entropy constrained quantization* since, in comparison to fixed rate SQ, a new optimization principle is introduced. This optimization criterion is to minimize the quantization error $D(x, Q(x))$ (3.4) subject to the constraint that the entropy does not exceed a certain value H_0,

$$H(Q(x)) \leq H_0 \tag{3.26}$$

In [Kos63] and [Zad66], it was shown for the assumption of high bit rates that the optimal quantizer for entropy constrained quantization is a uniform quantizer. The same result is described in [GP68] where it is shown analytically that the offset with respect to the 6-dB-per-bit curve for entropy constrained SQ of i.i.d. Gaussian random variables is only

$$\Delta_{\mathrm{SNR,vr}}|_{\mathrm{dB}} = 1.53 \text{ dB} \tag{3.27}$$

for high bit rates (also derived numerically in [GH67]). For lower bit rates, iterative algorithms to design entropy constrained quantizers similar to the LMQ have been proposed in, e.g., [Woo69], [Ber72], [FL84] and [NZ78] for the squared error distortion.
For the computation of the minimum average bit rate (3.25), it is assumed that a lossless binary coder exists so that the average bit rate reaches the entropy. The widest spread lossless binary coders are based on the Huffman [Huf52] or the Fano code [Fan61] which reach entropy only in few cases [CT91]. An alternative approach for lossless coding is Arithmetic Coding [Ris76] which has benefits compared to Huffman or Fano coders for very low bit rates and is therefore often used in applications for coding of images. Lossless entropy coding is investigated in various textbooks, e.g, [Mac03].

3.1.3 Intermediate Summary

Results related to the different SQ approaches are illustrated in Figure 3.4. In that figure, SNR curves based on the operational rate distortion functions for a memoryless Gaussian source for the LMQ and the entropy constraint SQ are shown as the dashed line with circle markers and the dotted line with asterisk markers, respectively. In addition, the curves for the high rate approximations for source optimized and A-Law SQ and for the rate distortion function are shown as the slash-dot-dotted line with triangle markers, the dashed line with x markers, and the solid line with square markers, respectively.
The SNR for entropy constrained SQ is the closest to the maximum achievable SNR according to the rate distortion function ($\Delta_{\mathrm{SNR,vr}} = 1.53$ dB). The asymptotic results for the LMQ are consistent with the PDF optimized SQ ($\Delta_{\mathrm{SNR,nu}} = 4.34$ dB), and the lowest SNR is achieved by the A-law SQ ($\Delta_{\mathrm{SNR,nu,A}} = 9.99$ dB). Nevertheless, the A-Law quantizer can be a reasonable choice in practical applications where signals are not necessarily stationary since its performance is independent from the PDF and the variance of the input signal in a wide dynamic range. The benefit

of variable rate SQ compared to the fixed rate approaches increases for PDFs with longer tails, e.g., the Laplacian or Gamma PDF [NZ78].
Even though entropy constrained SQ clearly outperforms fixed rate SQ, it is not always applicable in practice. The achievable bit rate is only an average value, which complicates a combination with a fixed rate digital transmission scheme. One solution is to buffer codewords, but in order to avoid that *buffer overflow* or *buffer exhaust* [Jel68] situations occur, the buffer sizes must be large which introduces additional delay. In addition to that, transmission bit errors may have fatal impact due to a strong error propagation.
Besides entropy constraint SQ, it will be shown in the next section that the quantization performance can also be increased by fixed rate VQ. We will see in Chapter 4 that with a Logarithmic Spherical Vector Quantizer, the performance of the entropy constraint SQ can be outperformed already for moderate vector dimensions which permits applications with low algorithmic delay. A fixed rate VQ also has the advantage that unequal error protection techniques can be developed to better combat transmission errors [Hei01][KSV06].

3.2 Vector Quantization (VQ)

In the previous section, it was shown that SQ does not reach the rate distortion function. Shannon [Sha48] stated that in order to reach the optimal performance, a

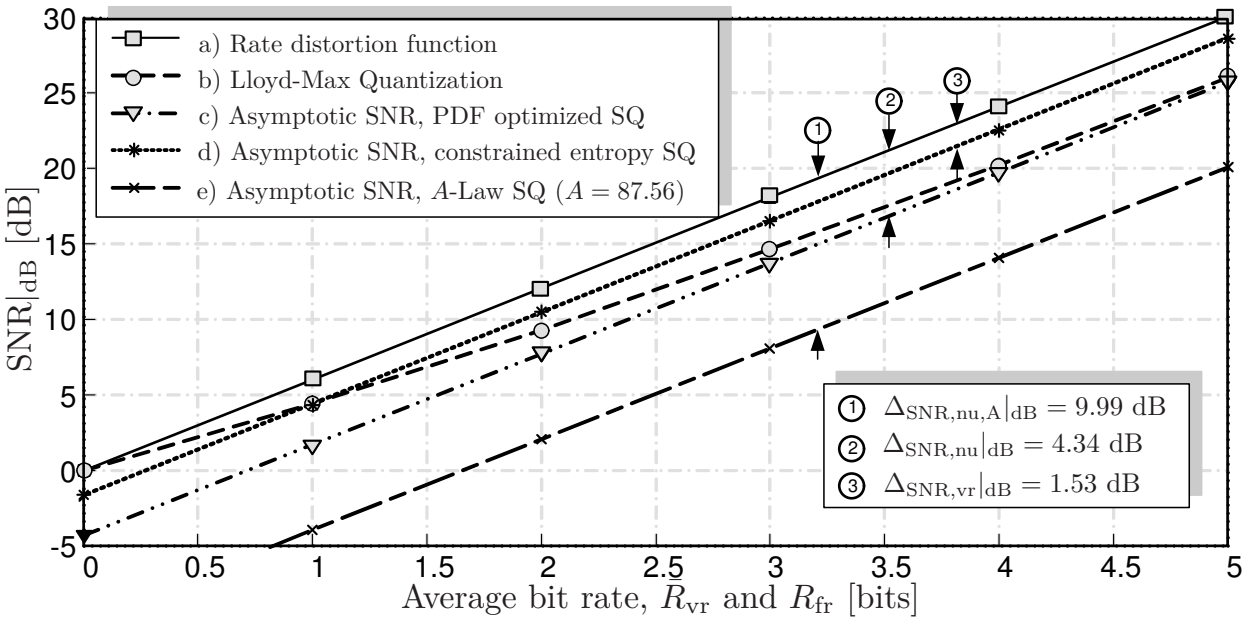


Figure 3.4: SNR over bit rate: a) acc. to rate distortion function, b) LMQ, c) asymptotic SNR for PDF optimized SQ, d) asymptotic SNR for constrained entropy SQ, and e) asymptotic SNR for A-Law SQ with $A = 87.56$ (G.711). The results are for a memoryless Gaussian random variable. The SNR offsets with respect to the rate distortion function are given as $\Delta_{\mathrm{SNR,nu}}|_{\mathrm{dB}}$, $\Delta_{\mathrm{SNR,nu,A}}|_{\mathrm{dB}}$, and $\Delta_{\mathrm{SNR,vr}}|_{\mathrm{dB}}$ according to [GP68]. The SNR values for the LMQ are taken from [JN84].

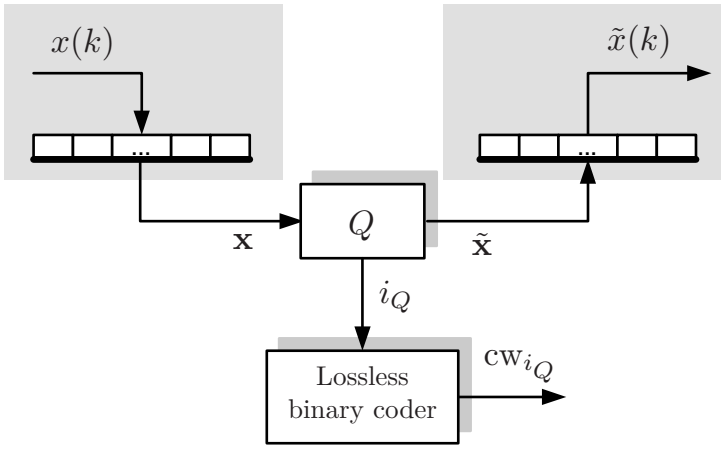


Figure 3.5: Principle of VQ.

quantizer may have to be designed for blocks of infinite length. In order to approach the rate distortion function more closely, in VQ, blocks of signal amplitudes $x(k)$ are collected and quantized jointly. Assuming stationary and ergodic signals, the blocks of sequential samples of the input signals $x(k)$ can be interpreted as sets of random variables in analogy to Section 2.1.
Even though it appears to make no sense to jointly quantize sets of i.i.d. random variables at the first glance, it will be shown that VQ benefits from the so-called *vector quantization advantages* even in this case.
Most of the results presented in the following can be interpreted as a generalization of SQ or, in other words, SQ is a special variant of VQ for vector dimension one.
The principle of VQ is shown in Figure 3.5. The input signal $x := x(k)$ with (quasi) continuous amplitude is buffered in a buffer of length L_v which is the dimension of the vector quantizer[1]. If the buffer is filled at time index k, the vector

$$\mathbf{x} = \begin{bmatrix} x(k-(L_v-1)) & \dots & x(k-1) & x(k) \end{bmatrix}^T \in \mathcal{X} \tag{3.28}$$

is fed into the quantizer Q to produce the representative

$$\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{x}(k-(L_v-1)) & \dots & \tilde{x}(k-1) & \tilde{x}(k) \end{bmatrix}^T = Q(\mathbf{x}) \in \tilde{\mathcal{X}}. \tag{3.29}$$

In the quantization procedure, from the finite alphabet $\tilde{\mathcal{X}}$, also denoted as the *vector codebook*, that *codevector* $\tilde{\mathbf{x}}_{i_{Q,\mathrm{vq}}}$ with index $i_{Q,\mathrm{vq}}$ is determined which reconstructs input vector $\mathbf{x}$ from the infinite alphabet $\mathcal{X}$ with the minimum quantization cost given a certain distortion function, e.g.,

$$i_{Q,\mathrm{vq}} = \underset{0 \le i_{\mathrm{vq}} < N_{\mathrm{vq}}}{\arg\min} \; \| \mathbf{x} - \tilde{\mathbf{x}}_{i_{\mathrm{vq}}} \|^2 \;\; \forall \; \tilde{\mathbf{x}}_{i_{\mathrm{vq}}} \in \tilde{\mathcal{X}} \tag{3.30}$$

based on the squared error and with N_{vq} as the number of vectors in the vector codebook $\tilde{\mathcal{X}}$. This quantization procedure is denoted also as the *nearest neighbor*

[1]Note that a new parameter for the vector dimension, L_v, is introduced here compared to parameter N as defined in Chapter 2.1 to distinguish between rate distortion and asymptotic quantization theory.

codevector search. The output vector of the quantizer is stored in the output buffer to produce the output signal $\tilde{x} := \tilde{x}(k)$. In analogy to Figure 3.3, the codevector index i_Q may be transformed into a codeword in a lossless binary coder. The concept of entropy constrained VQ [LZ94], however, will not be considered here due to the reasons given in Section 3.1.3 so that all codewords have the same bit width $R_{\mathrm{vq}} = \log_2(N_{\mathrm{vq}})$. In analogy to (2.27) and (2.6), the mean of the quantization cost function per vector coordinate for a squared error criterion (MSE) is specified as the **per vector coordinate distortion**

$$D = \frac{E\{\| \mathbf{x} - Q(\mathbf{x}) \|^2\}}{L_v} \tag{3.31}$$

which can be transformed also into a **per vector distortion**

$$D^* = D \cdot L_v = E\{\| \mathbf{x} - Q(\mathbf{x}) \|^2\}. \tag{3.32}$$

The *quantization cells* are the analogon to the quantization intervals in SQ (3.2) and, with respect to the squared error criterion, defined as

$$C_{\mathrm{vq},\tilde{\mathbf{x}}} := \{\mathbf{x} \in \mathcal{X} : \| \mathbf{x} - \tilde{\mathbf{x}} \|^2 \leq \| \mathbf{x} - \tilde{\mathbf{x}}' \|^2 \ \ \forall\, \tilde{\mathbf{x}}' \neq \tilde{\mathbf{x}}, \text{ and } \tilde{\mathbf{x}}', \tilde{\mathbf{x}} \in \tilde{\mathcal{X}}\} \tag{3.33}$$

with the codevectors located in the center of gravity of the cell. It is shown in [Zad66] and [Ger79] that (3.15) can be generalized to compute the minimum achievable quantization distortion of a source optimized fixed rate VQ of dimension L_v for the assumption of high bit rates as

$$D(N_{Q,L_v}) = C_{\mathrm{q}}(L_v) \cdot (N_{Q,L_v})^{-2/L_v} \cdot \| p(\mathbf{x}) \|_{L_v/(L_v+2)} \tag{3.34}$$

with

$$\| p(\mathbf{x}) \|_{L_v/(L_v+2)} = \left(\int_{\mathcal{X}} (p(\mathbf{x}))^{L_v/(L_v+2)} \right)^{(L_v+2)/L_v}. \tag{3.35}$$

In this equation, $p(\mathbf{x}) := p_{\mathbf{x}}(\mathbf{x})$ is the multivariate PDF of the input signal vector $\mathbf{x}$, $N_{Q,L_v} := N_{\mathrm{vq}}$ is the number of codevectors in vector codebook $\tilde{\mathcal{X}}$, written in a notation to highlight the analogy to the number of quantization reconstruction levels $N_{Q,1} := N_Q$ from Section 3.1 for vector dimension L_v, and $C_{\mathrm{q}}(L_v)$ is a constant to be discussed in Section 3.2.1. In analogy to (3.14), necessary condition to achieve this minimum distortion is that the normalized quantizer codevector density function is

$$\lambda(\mathbf{x}) = \frac{(p(\mathbf{x}))^{L_v/(L_v+2)}}{\int\limits_{\mathcal{X}} (p(\mathbf{y}))^{L_v/(L_v+2)} \mathrm{d}\mathbf{y}} \tag{3.36}$$

and a (quasi) continuous function due to the assumption of high bit rates.

3.2.1 The VQ Advantages

According to (3.34), the term for the determination of the overall distortion is composed of three independent parts, the constant $C_{\mathrm{q}}(L_v)$, the part related to the number N_{Q,L_v} of codevectors, and the part that depends on the shape of the multivariate PDF, $\| p(\mathbf{x}) \|_{L_v/(L_v+2)}$. In [LG89], the performance of VQ is compared to that of SQ based on the same effective bit rate per vector coordinate

$$R_{\text{eff, vq}} = \frac{\log_2(N_{Q,L_v})}{L_v} \stackrel{!}{=} \log_2(N_{Q,1}) \tag{3.37}$$

with $N_{Q,1} = N_Q$ from Section 3.1 since SQ is a special variant of VQ for $L_v = 1$. The *vector quantizer advantage* is defined as the distortion achievable by source optimized SQ in relation to the distortion achievable by source optimized VQ

$$\frac{D_{L_v=1}(N_{Q,1})}{D_{L_v}(N_{Q,L_v})} = \underbrace{\frac{C_{\mathrm{q}}(1)}{C_{\mathrm{q}}(L_v)}}_{F(L_v)} \cdot \underbrace{\frac{N_{Q,1}^{-2}}{N_{Q,L_v}^{-2/L_v}}}_{1} \cdot \underbrace{\frac{\| \hat{p}(x) \|_{1/3}}{\| p(\mathbf{x}) \|_{L_v/(L_v+2)}}}_{S(L_v)\cdot M(L_v)} \tag{3.38}$$

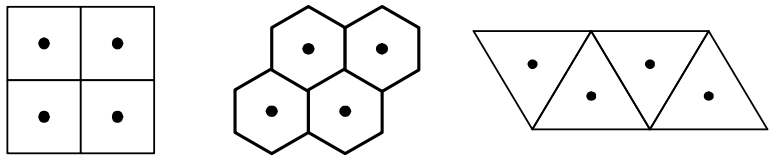
and can be grouped into the *Space Filling Advantage* $F(L_v)$, the *Shape Advantage* $S(L_v)$, and the *Memory Advantage* $M(L_v)$. Each of these advantages contributes to a lower distortion achievable in VQ compared to SQ and will be explained in the following.

3.2.1.1 The Space Filling Advantage

According to (3.38), $F(L_v)$ is a function of the constants $C_{\mathrm{q}}(L_v)$ and $C_{\mathrm{q}}(1)$ and independent of the PDF of the input signal. To better understand the role of the constant $C_{\mathrm{q}}(L_v)$, the special case of a uniform PDF $p(\mathbf{x})$ is considered. In that case, the overall distortion (3.34) depends only on the constant $C_{\mathrm{q}}(L_v)$ since

$$\| p(\mathbf{x}) \|_{L_v/(L_v+2)} = 1 \qquad \Leftrightarrow \qquad p(\mathbf{x}) \text{ uniform} \tag{3.39}$$

and hence on the filling of the L_v-dimensional vector space by quantization cells. In general, each quantization cell is bounded by $(L_v - 1)$-dimensional hyperplanes and also called a (convex) *polytope* P or *Voronoi Region*. In [Ger79], it is conjectured that for a specific dimension L_v, the best fitting of quantization cells is achieved if all quantization cells are congruent (except for those at the boundaries of the vector space) which is called a *tessellation*. Effectively, however, not all quantization cell shapes are suitable to generate a tessellation. Considering for example $L_v = 2$, three types of quantization cell shapes that form a tessellation are the rectangle, the regular hexagon, and the equilateral triangle as demonstrated in Figure 3.6 a), b) and c), respectively. The group of possible cell shapes and hence polytopes to generate a tessellation in L_v dimensions are called *admissible polytopes* $\mathcal{P}_{L_v}$. In order to assess the suitability of different admissible polytopes $P \in \mathcal{P}_{L_v}$ for



a) Rectangle b) Hexagon c) Triangle

Figure 3.6: Admissible polytopes for $L_v = 2$

quantization, in [Ger79] the so-called *normalized inertia*

$$\mathrm{NI}(P) = \frac{1}{L_v} \cdot \frac{\int\limits_P \| \mathbf{x} - \tilde{\mathbf{x}} \|^2 \, \mathrm{d}\mathbf{x}}{(V(P))^{(L_v+2)/L_v}} \tag{3.40}$$

is introduced, with $V(P)$ as the L_v-dimensional volume of polytope P. Considering all admissible polytopes for a given dimension L_v, the constant $C_\mathrm{q}(L_v)$ is

$$C_\mathrm{q}(L_v) = \inf_{P \in \mathcal{P}_{L_v}} \mathrm{NI}(P), \tag{3.41}$$

for the *optimal polytope* in L_v dimensions. According to [Ger79], an optimal polytope exists for each dimension L_v but, unfortunately, little is known about optimal polytopes for $L_v > 3$. Closed form solutions are given for $L_v = 1$ as known from scalar quantization, $C_\mathrm{q}(1) = \frac{1}{12}$, for $L_v = 2$ with the hexagon as the optimal polytope [Tot59b] and $C_\mathrm{q}(2) = 5\sqrt{3}/108$, and for $L_v = 3$ with the *regular truncated octahedron* as the optimal polytope [CS93] with $C_\mathrm{q}(3) = 0.078543$.
In the literature, bounds are given for $C_\mathrm{q}(L_v)$, namely the *Sphere Lower Bound* based on the fact that every convex polytope has a moment that is greater than that of a sphere, and the *Conway and Sloan Conjectured Lower Bound* [CS85]. The measured results for lattice quantizers in different dimensions (refer to Section B.1 in the *supplement document* [Krü09]) are given as the *Lattice Upper Bound*.
The space filling advantage $F(L_v)$ in dB related to the Sphere Lower, the Conway and Sloane Lower, and the Lattice Upper Bound are listed, e.g., in [Erd04] and [LG89]. All bounds asymptotically reach

$$\lim_{L_v \to \infty} 10 \cdot \log_{10}(F(L_v)) = 1.53 \text{ dB}. \tag{3.42}$$

3.2.1.2 The Shape Advantage

The shape advantage $S(L_v)$ in (3.38) depends on the multivariate PDF of the input signal. For memoryless (uncorrelated) sources, no memory advantage can be exploited and VQ benefits only from the shape advantage which is defined as

$$S(L_v) = \frac{\| \hat{p}(x) \|_{1/3}}{\| p(\mathbf{x}) \|_{L_v/(L_v+2)}}. \tag{3.43}$$

The multivariate PDF $p(\mathbf{x})$ for a memoryless source can be computed from the one-dimensional (marginal) PDF (see the example in Figure 3.7), denoted as $\hat{p}(x)$, as

$$p(\mathbf{x}) = \prod_{i=0}^{L_v-1} \hat{p}(x). \tag{3.44}$$

If the signal to be quantized has a uniform distribution, the shape advantage is

$$S_U(L_v) = 1. \tag{3.45}$$

Assuming an i.i.d. Gaussian distributed source, the shape advantage is [LG89]

$$S_G(L_v) = \frac{3^{3/2}}{\left(\frac{L_v+2}{L_v}\right)^{(L_v+2)/2}}. \tag{3.46}$$

with the asymptotic value for infinite dimensions

$$\lim_{L_v\to\infty} 10 \cdot \log_{10}(S_G(L_v)) = 2.81 \text{ dB} \tag{3.47}$$

Values for the shape advantage in dB for Gaussian, Laplacian and Gamma PDFs are, e.g., given in [Erd04] and [LG89].

3.2.1.3 The Memory Advantage

Given sources with memory, in contrast to a PDF optimized SQ, a PDF optimized VQ benefits from the correlation immanent to the input signal, denoted as $M(L_v)$ in (3.38). Note that in the following only the case of linear dependencies is considered, an example for non-linear dependencies is described in [MSG85].
In analogy to the results from Section 2.3, the memory advantage shall be explained based on the example of a correlated Gaussian random variable with zero mean. The multivariate PDF is

$$p_{\mathbf{X}}(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mathbf{0}, \Phi_{\mathbf{X}}) = \frac{1}{(2\pi)^{N/2}} \cdot \frac{1}{\sqrt{\det(\Phi_{\mathbf{X}})}} \cdot \exp(-\frac{1}{2}\mathbf{x} \cdot \Phi_{\mathbf{X}}^{-1} \cdot \mathbf{x}^T) \tag{3.48}$$

with the covariance matrix $\Phi_{\mathbf{X}}$ and the determinant thereof, $\det(\Phi_{\mathbf{X}})$, as introduced in (2.41). An example multivariate PDF is illustrated by the three-dimensional plot in Figure 3.7 a) for $L_v = 2$ and a covariance matrix of

$$\Phi_{\mathbf{X}} = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho \\ \sigma^2 \cdot \rho & \sigma^2 \end{bmatrix} = \sigma^2 \cdot \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \tag{3.49}$$

with $\sigma^2 = 1.0$ and $\rho = 0.75$. Sequences of samples are stored in the vector $\mathbf{x} = \begin{bmatrix} x(k-1) & x(k) \end{bmatrix}^T$. From (3.38), combined shape and memory advantage are given

as

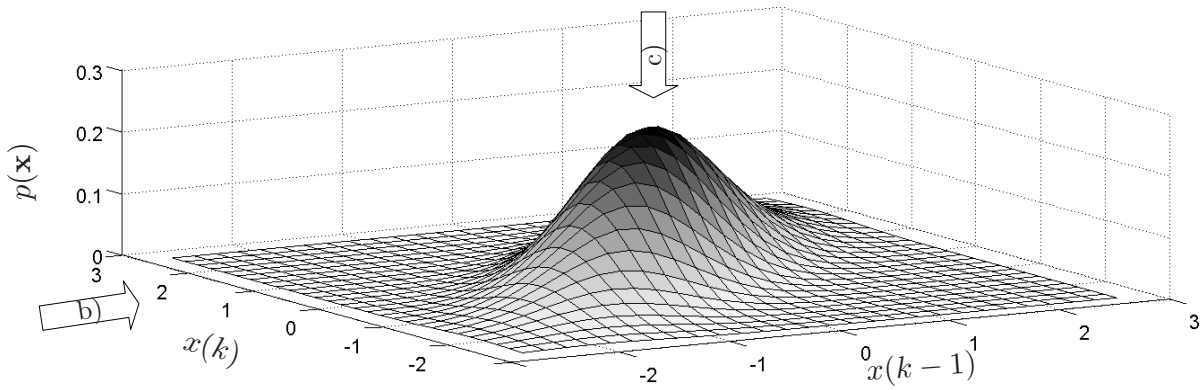
$$S(L_v) \cdot M(L_v) = \frac{\| \hat{p}(\mathbf{x}) \|_{1/3}}{\| p(\mathbf{x}) \|_{L_v/(L_v+2)}} \tag{3.50}$$

In order to separate both parts, the joint PDF $p'(\mathbf{x})$ related to L_v independent variables with marginal PDF $\hat{p}(\mathbf{x})$ is computed according to (3.44) as

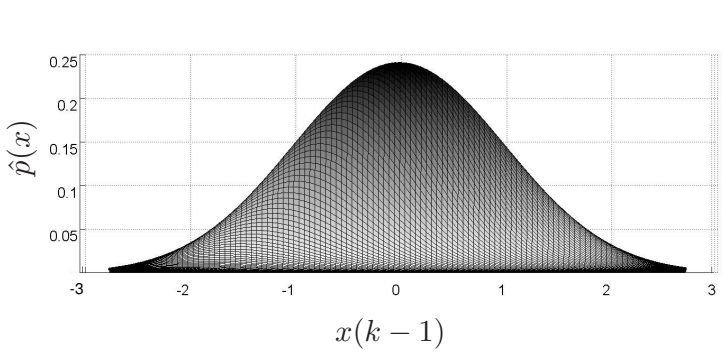
$$p'(\mathbf{x}) = \prod_{i=0}^{L_v-1} \hat{p}(x) \tag{3.51}$$

to extend (3.50),

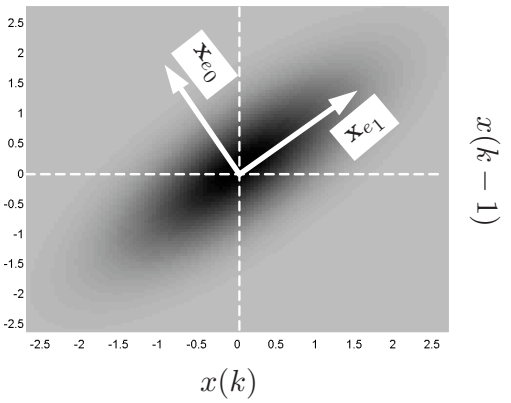
$$S(L_v) \cdot M(L_v) = \underbrace{\frac{\| \hat{p}(\mathbf{x}) \|_{1/3}}{\| p'(\mathbf{x}) \|_{L_v/(L_v+2)}}}_{S(L_v)} \cdot \underbrace{\frac{\| p'(\mathbf{x}) \|_{L_v/(L_v+2)}}{\| p(\mathbf{x}) \|_{L_v/(L_v+2)}}}_{M(L_v)} . \tag{3.52}$$



a) Three-dimensional plot of the PDF for $L_v = 2$



b) Marginal PDF $\hat{p}(x(k))$



c) Top view of three dimensional PDF.

Figure 3.7: PDF for Gaussian source with memory (covariance matrix $\Phi_{\mathbf{X}}$ and $\rho = 0.75$). In a), the three-dimensional plot of the PDF is shown. In b), the marginal PDF as viewed from the position marked by the respective arrow in a) is depicted. In c), the top view of the PDF from the view position according to the corresponding arrow in a) is presented together with the Eigenvectors $\mathbf{x}_{e_0}$ and $\mathbf{x}_{e_1}$.

The first part of this equation is identical to the shape advantage $S(L_v)$ in (3.43) (since $p'(\mathbf{x})$ in (3.51) is identical to $p(\mathbf{x})$ in (3.44)). Consequently the memory advantage is

$$M(L_v) = \frac{\parallel p'(\mathbf{x}) \parallel_{L_v/(L_v+2)}}{\parallel p(\mathbf{x}) \parallel_{L_v/(L_v+2)}} \tag{3.53}$$

In [LG89], it is shown that for correlated Gaussian random variables with covariance matrix $\Phi_{\mathbf{X}}$ in general

$$\parallel p(\mathbf{x}) \parallel_{L_v/(L_v+2)} = 2\pi \cdot \left(\frac{L_v+2}{2}\right)^{(L_v+2)/2} \cdot \det(\Phi_{\mathbf{X}})^{1/L_v}. \tag{3.54}$$

Considering the numerator part in (3.53), the corresponding covariance matrix is

$$\Phi'_{\mathbf{X}} = \begin{bmatrix} \hat{\sigma}^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \hat{\sigma}^2 \end{bmatrix} = \hat{\sigma}^2 \cdot \mathbf{I}_{L_v} \tag{3.55}$$

with $\hat{\sigma}^2$ as the variance related to the marginal PDF $\hat{p}(x)$. Accordingly, the norm (3.54) is

$$\parallel p'(\mathbf{x}) \parallel_{L_v/(L_v+2)} = 2\pi \cdot \left(\frac{L_v+2}{2}\right)^{(L_v+2)/2} \cdot \hat{\sigma}^2. \tag{3.56}$$

Considering the computation of the norm for $p(\mathbf{x})$, in analogy to (2.43), $\Phi_{\mathbf{X}}$ can be diagonalized with respect to its principal axes without changing the norm. Given the Eigenvalues $\lambda_{\mathbf{X},i}$ as the result of the matrix diagonalization, the memory advantage can hence be computed as

$$M(L_v) = \frac{\hat{\sigma}^2}{\left(\prod_{i=0}^{(L_v-1)} \lambda_{\mathbf{X},i}\right)^{1/L_v}} = \frac{\frac{1}{L_v} \cdot \left(\sum_{i=0}^{(L_v-1)} \lambda_{\mathbf{X},i}\right)}{\exp\left(\frac{1}{L_v} \cdot \sum_{i=0}^{(L_v-1)} \ln(\lambda_{\mathbf{X},i})\right)}. \tag{3.57}$$

with the variance $\hat{\sigma}^2$ related to the marginal PDF computed from the Eigenvalues of the covariance matrix in analogy to (2.36). In the example in Figure 3.7, the Eigenvectors for the matrix diagonalization are shown as $\mathbf{x}_{e_0}$ and $\mathbf{x}_{e_1}$ in part c), the view from above.

3.2.2 Asymptotic VQ Performance

Given a memoryless Gaussian source, in comparison to SQ, VQ benefits from the space filling and the shape advantage. Assuming high bit rates, given the asymptotic offset related to source optimized SQ from (3.17) and the asymptotic values for the

VQ advantages in (3.42) and (3.47), the asymptotic logarithmic SNR related to VQ with infinite block lengths is equal to the rate distortion function,

$$\begin{aligned}\underset{L_V\to\infty}{\mathrm{SNR}_{\mathrm{vq,G}}(R_{\mathrm{eff,vq}})|_{\mathrm{dB}}} &= \mathrm{SNR}_{\mathrm{sq,nu,G}}(R_{\mathrm{fr}})|_{\mathrm{dB}} + \underbrace{10\log_{10}(F(\infty))}_{1.53\mathrm{dB}} + \underbrace{10\log_{10}(S_G(\infty))}_{2.81\mathrm{dB}} \\ &= 6.02\ \mathrm{dB}\cdot R_{\mathrm{eff,vq}}, \end{aligned} \tag{3.58}$$

with $R_{\mathrm{fr}} = R_{\mathrm{eff,vq}}$. In addition to that, by comparing (2.45) and (3.57), it is obvious that source optimized VQ asymptotically also reaches the rate distortion function for correlated Gaussian sources.

3.2.3 VQ Design for Low Bit Rates

The results for VQ so far were based on the assumption of high bit rates. For lower bit rates, the generalization of the LMQ (Section 3.1.1.4) for vectors is described in [LBG80]. According to that paper, a VQ codebook is the output of a "training" procedure for a large amount of training data. The resulting VQ consists of the codevectors which must be stored in a lookup table. This type of quantizer has a high performance even for low bit-rates and benefits from all VQ advantages. In practical applications, however, trained quantizers are impractical due to high memory requirements to store all codevectors. In addition to that, because the codebook is *unstructured*, a fast procedure to find the nearest neighbor for a signal vector to be quantized is not straight forward [Vid86][MM84]. Nevertheless, trained quantizers are often used for specialized applications for low vector dimensions or bit rates, e.g., in the quantization of the spectral envelope in speech coding, [PA93][ETS00][ETS01].

3.2.4 VQ Application Examples

In Section B in the *supplement document* [Krü09] two example concepts for VQ are reviewed which are of high importance in a lot of source coding applications. Due to the employment of *structured* codebooks, highly efficient nearest neighbor quantization procedures can be realized with low computational complexity and memory consumption and therefore will be the basis for two candidate VQ realizations in Section 4.

3.3 Discussion

In the first part of this chapter, different approaches for SQ were briefly summarized. It was shown that for high bit rates, source optimized SQ can be realized as a combination of a compressor function and a uniform quantizer. The compressor function was generalized to a point density function, and the optimal point density can be computed from the PDF of the signal to be quantized. Also, it was described that the highest quantization performance can be achieved by entropy constrained

SQ. However, for the target application low delay audio coding, a variable bit rate is not useful. Required extensive buffering and a very high sensitivity against transmission bit errors make this approach impractical.
In the second part of this chapter, fixed rate VQ was reviewed as a generalization of SQ. It was shown that for high bit rates, a source optimized VQ adapts its codevector point density according to the PDF of the input signal and, compared to SQ, benefits from the VQ advantages. Source optimized VQ asymptotically reaches the rate distortion function for infinite block lengths.

4

Logarithmic Spherical VQ (LSVQ)

In the previous section, it was shown in (3.36) that the density of codevectors for a source optimized VQ can be derived from the multivariate PDF of the input signal for high bit rates. Unfortunately, this knowledge does not provide any information on how to construct a source optimized VQ that is also well applicable in practice. In all previous chapters, the signal to be quantized was assumed to be stationary and the PDF well known. Targeting the quantization of real audio signals, identifying a PDF that accurately describes all parts of an input signal is not straight forward: Audio signals are not stationary, can have a very high dynamic and may also have an arbitrary PDF. An approach known from speech coding is to consider short segments of speech signals which are approximately stationary [VM06]. Each segment, however, may have a different characteristic so that a VQ should be adapted to each single segment independently. Since VQs are mostly designed offline, designing a new VQ for each segment is certainly not an option. In most applications, therefore, fixed quantizers designed for memoryless sources are combined with common techniques such as Linear Prediction (LP) or Transform Coding (TC) to exploit linear correlation. LP is also suitable for low delay audio coding and will be studied in detail in Chapter 5.

Designing an optimal fixed VQ for memoryless sources still is a complex task since the dynamics as well as the PDF of the memoryless signal to be quantized are unknown. In the early days of speech coding, measurements were proceeded to find a PDF that is common for speech, e.g., [Ric73], [Nol74], [Ata82][BJW74]. In the end it turned out that a Gaussian distribution is a good approximation of the distribution of normalized LP residual signals for speech.

Considering audio signals, the assumption of a common PDF does not make sense. Instead, a quantizer should be optimized to have reasonable performance for all types of signals. In this context, the assumption of a Gaussian distribution as a

worst case scenario is useful since it has the highest differential entropy [Ber71] and therefore is the most "difficult" source distribution for quantization.

4.1 Motivation for Spherical VQ (SVQ)

Given the multivariate PDF of a memoryless i.i.d. Gaussian source with variance σ^2 and zero mean, it can be shown that the optimal normalized codevector density of dimension L_v for high bit rates according to (3.36) is

$$\lambda_G(\mathbf{x}) = C \cdot \exp(-\frac{L_v}{L_v+2} \cdot \frac{1}{2\sigma^2} \cdot \mathbf{x}^T \cdot \mathbf{x}) \quad \forall \quad \mathbf{x} \in \mathbb{R}^{L_v}. \tag{4.1}$$

with the unknown constant C. This qualitative result shall not be discussed here in detail. However, since $\lambda_G(\mathbf{x})$ is constant for vectors $\mathbf{x}$ with equal absolute value ($\|\mathbf{x}\| = \text{const}$), (4.1) is a good argument to arrange all codevectors uniformly on shells of spheres.
One way to achieve that all codevectors are arranged on shells of spheres is known from speech coding as gain-shape VQ which was proposed in [SG84]. In one of the first realizations of a gain-shape VQ in [AS84], Gaussian distributed sequences (the shape vectors) are multiplied with an amplitude (the gain factor) to generate the innovation part of a speech signal. Later, the Gaussian sequences were replaced by multi-pulses [Ata86] or ternary pulses known as *algebraic codebooks* [SS87], [XIB88], [IX89], [SA89], [Sal89]. Still today, this principle is the basis for most state-of-the-art standardized speech codecs like the ITU-T G.729 codec [ITU96] and the Adaptive Multirate Narrowband (AMR) and Wideband (AMR-WB) Speech Codecs [ETS00], [ETS01].
The algebraic codebooks were designed for very low bit rates. However, it will be shown in the course of this chapter that in order to achieve higher quantization performance at higher bit rates, algebraic codebooks are no longer suitable. Instead, new VQ concepts must be developed. In this context, Spherical VQ (SVQ) turns out to be a promising technique and will therefore be studied in detail.
Most often, realizations of SVQ are based on codevectors which are constructed based on so-called spherical codes. In this context, the spherical code can be considered as a rule to generate codevectors on the surface of spheres. In the literature, remarkable results on the analysis of spherical codes in general are in particular given in [Ham96], [HZ97a], [HZ97b]. In [HZ03], the same authors propose SVQ based on a so-called wrapped spherical code in combination with a source optimized SQ (see Section 3.1.1.4) in gain-shape fashion for the coding of Gaussian sources. In contrast to this, the concept of Logarithmic SVQ (LSVQ) is based on a combination of SVQ with **logarithmic** SQ (see Section 3.1.1.5). The employment of a logarithmic rather than a source optimized SQ is motivated by the fact that audio signals reveal a wide dynamic range and have an unknown PDF in practice.

In the first part of this chapter, the concept of LSVQ is defined and novel theoretical

results are presented. It is shown for high bit rate assumptions that LSVQ approximately achieves a constant quantization SNR independent of the PDF of the input signal over a wide dynamic range. Then, very similar to the *Sphere Lower Bound* for VQ in general ([Ger79], Section 3.2.1.1), a new lower bound for the achievable quantization distortion related to SVQ and LSVQ and a high rate estimate for the theoretical performance of LSVQ are derived. Integral part of the derivation of the high rate estimate is the calculation of the optimal allocation of bit rate for the gain and shape components given a fixed overall bit budget.
In the second part of the chapter, three exemplary SVQ realizations will be introduced. The presented examples are designed for high performance quantization, but at the same time, low computational cost and memory consumption is preserved due to newly developed nearest neighbor quantization procedures. In order to assess the achievable quality, the quantization performance of all quantizers measured for stationary memoryless i.i.d. Gaussian input signals will be compared to the theoretical results from the first part at the end of this chapter.

4.2 Theory of LSVQ

In the literature, most of the work on spherical codes are mathematical reviews on the asymptotic code density, e.g., [Ast84], [Tot59a], and [Cox68]. A good summary of a lot of aspects related to spherical codes is given in [Ham96]. In contrast to this, the purpose of this chapter is mainly to discuss aspects related to the application of spherical codes for LSVQ (refer to [KSGV08] also).

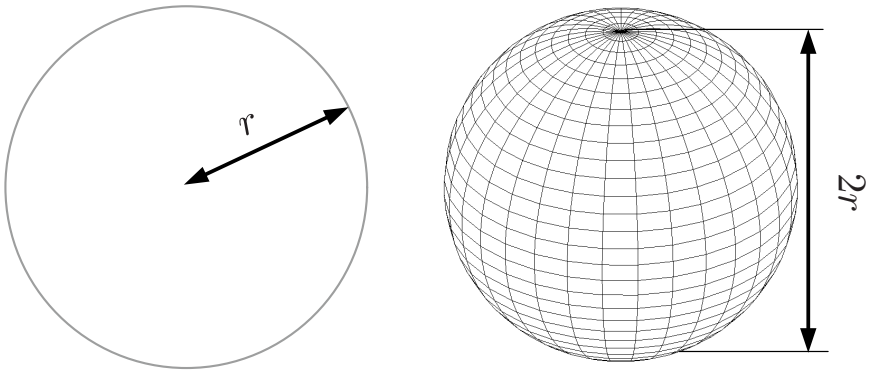
4.2.1 Properties of Spheres

Some general properties of spheres shall be introduced first: An L_v-dimensional sphere with radius r, denoted as $\mathcal{S}_{L_v}^{(r)}$, is defined as the amount of all vectors $\mathbf{x}$ in L_v-dimensional space with a distance r to the origin,

$$\mathcal{S}_{L_v}^{(r)} := \{\mathbf{x} = \begin{bmatrix} x_0 & \dots & x_{L_v-1} \end{bmatrix}^T \in \mathbb{R}^{L_v} : \|\mathbf{x}\| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} = r\}. \tag{4.2}$$

Examples are well known for $L_v = 2$ and $L_v = 3$ as a circle and a ball as shown in Figure 4.1 a) and b), respectively. An L_v-dimensional sphere with a radius $r = 1.0$ is called a *unit sphere*, $\mathcal{S}_{L_v}^{(1.0)}$. The content within the unit sphere [Cox73] (denoted as the *volume* of a sphere) is

$$\begin{aligned} V_{\mathcal{S}_{L_v}}^{(1.0)} &= \int_{\|\mathbf{x}\| \le 1} \mathrm{d}\mathbf{x} = \frac{2 \cdot \pi^{L_v/2}}{L_v \cdot \Gamma(L_v/2)} = \frac{\pi^{L_v/2}}{\Gamma((L_v+2)/2)} \\ &= \begin{cases} \frac{\pi^{L_v/2}}{(L_v/2)!} & \text{if } L_v \text{ even} \\ \frac{2^{L_v}\,\pi^{(L-v-1)/2}((L_v-1)/2)!}{L_v!} & \text{if } L_v \text{ odd} \end{cases} \end{aligned} \tag{4.3}$$



a) Sphere for $L_v = 2$ b) Sphere for $L_v = 3$

Figure 4.1: Examples for spheres for a) $L_v = 2$ and b) $L_v = 3$.

with the Gamma function $\Gamma(y) = \int_0^\infty e^{-t} \cdot t^{y-1} \mathrm{d}t$ [BS91]. The volume of an L_v-dimensional sphere with radius r is calculated from the volume of the unit sphere as

$$V_{\mathcal{S}_{L_v}}^{(r)} = V_{\mathcal{S}_{L_v}}^{(1.0)} \cdot r^{L_v}. \tag{4.4}$$

The *shell of a sphere* is defined as all points which are located on the surface of the sphere and hence have the same distance r to the origin. The surface area content is given as

$$S_{\mathcal{S}_{L_v}}^{(r)} = \int_{\mathbf{x} \in \mathcal{S}_{L_v}^{(r)}} \mathrm{d}\mathbf{x} = V_{\mathcal{S}_{L_v}}^{(1.0)} \cdot L_v \cdot r^{(L_v - 1)}. \tag{4.5}$$

An interesting fact is that the volume of the unit sphere as a function of dimension L_v grows for increasing dimension L_v, then reaches a maximum value for $L_v = 5$ and asymptotically reaches $\lim_{L_v \to \infty} V_{\mathcal{S}_{L_v}}^{(1.0)} = 0$ as demonstrated by Figure 4.2.

4.2.2 Definition of LSVQ

LSVQ is a special type of fixed rate gain-shape VQ in which each input vector $\mathbf{x}$ is decomposed into its absolute value and a normalized version with unit absolute value, referred to as the *gain* and the *shape* components

$$g = \|\mathbf{x}\| = \sqrt{\mathbf{x}^T \cdot \mathbf{x}} \qquad \text{and} \qquad \mathbf{c} = \frac{1}{g} \cdot \mathbf{x} = \frac{\mathbf{x}}{\sqrt{\mathbf{x}^T \cdot \mathbf{x}}}. \tag{4.6}$$

Both components are quantized by the two quantizers

$$Q_g : g \mapsto \tilde{g}, \qquad \tilde{g} \in \tilde{\mathcal{X}}_g \tag{4.7}$$

$$Q_{\text{svq}} : \mathbf{c} \mapsto \tilde{\mathbf{c}}, \qquad \tilde{\mathbf{c}} \in \tilde{\mathcal{X}}_{\text{svq}} \tag{4.8}$$

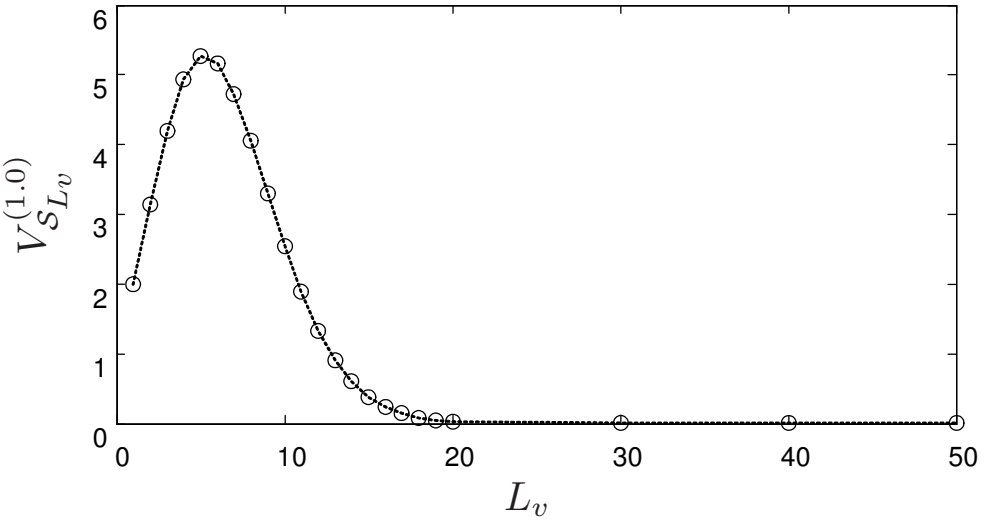


Figure 4.2: Volume of unit sphere over dimension L_v.

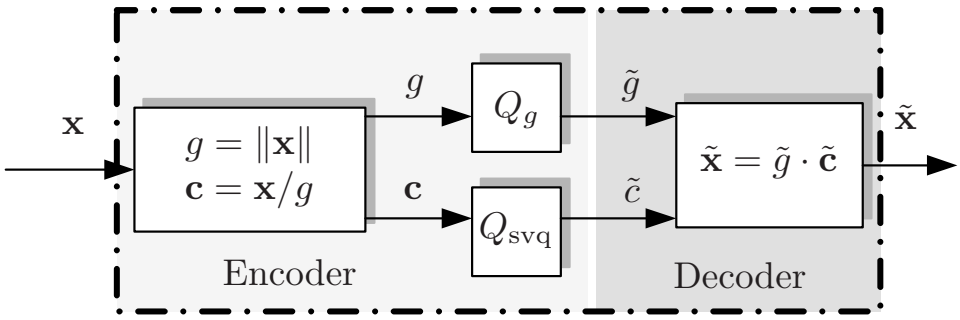


Figure 4.3: Parallel quantization of the gain and shape component in LSVQ.

with the codebooks $\tilde{\mathcal{X}}_g$ and $\tilde{\mathcal{X}}_{\text{svq}}$ for the SQ of the gain component and the VQ of the shape component, respectively. From the quantized shape and gain components, the quantized overall signal vector is reconstructed as

$$\tilde{\mathbf{x}} = \tilde{\mathbf{c}} \cdot \tilde{g}. \tag{4.9}$$

The overall LSVQ

$$Q_{\text{lsvq}} : \mathbf{x} \mapsto \tilde{\mathbf{x}}, \qquad \tilde{\mathbf{x}} \in \tilde{\mathcal{X}}_{\text{lsvq}} = \tilde{\mathcal{X}}_g \times \tilde{\mathcal{X}}_{\text{svq}} \tag{4.10}$$

is hence defined as the combination of both quantizers,

$$Q_{\text{lsvq}} = Q_g \bigcirc Q_{\text{svq}} \tag{4.11}$$

with $\tilde{\mathcal{X}}_{\text{lsvq}}$ as the *overall* codebook. The combination of Q_g and Q_{svq}, represented by the operator $\bigcirc$, is based on a parallel or independent approach as demonstrated by Figure 4.3. Variants of LSVQ based on a *sequential* and a *joint* approach for the combination of the two quantizers are discussed in detail in Section C of the *supplement document* [Krü09]. All codevectors in codebook $\tilde{\mathcal{X}}_{\text{svq}}$ have unit absolute value and are therefore located on the surface (shell) of a unit sphere of dimension

L_v. These codevectors will be denoted as the *spherical codevectors* in the following. Given the effective *bit rate per vector coordinate* for the two quantizers as $R_{\text{eff},g}$ and $R_{\text{eff,svq}}$, the number of quantization reconstruction levels of Q_g is $N_g = 2^{R_{\text{eff},g} \cdot L_v}$, and the number of spherical codevectors related to Q_{svq} is $N_{\text{svq}} = 2^{R_{\text{eff,svq}} \cdot L_v}$, respectively. The overall number of codevectors related to the combination of both quantizers, referred to as *overall codevectors*, is

$$N_{\text{lsvq}} = 2^{R_{\text{eff,lsvq}} \cdot L_v} = N_g \cdot N_{\text{svq}} = 2^{(R_{\text{eff,svq}} + R_{\text{eff},g}) \cdot L_v} \tag{4.12}$$

with the overall effective bit rate per vector coordinate $R_{\text{eff,lsvq}}$.
In analogy to Figure 3.1, the quantizer Q_g outputs an index $i_{Q,g} \in \{0 \ldots (N_g - 1)\}$, and the quantizer Q_{svq} an index $i_{Q,\text{svq}} \in \{0 \ldots (N_{\text{svq}} - 1)\}$, representing one quantization reconstruction level and spherical codevector, respectively. Following the parallel approach from Figure 4.3 for a squared error criterion (2.7), given an input signal vector $\mathbf{x}$ and its decomposition into the shape and the gain component, g and $\mathbf{c}$, the optimal quantization indices are determined in a *nearest neighbor* approach in analogy to (3.30),

$$i_{Q,\text{svq}} = \underset{0 \le i_{\text{svq}} < N_{\text{svq}}}{\arg\min} \; \|\mathbf{c} - \tilde{\mathbf{c}}_{i_{\text{svq}}}\|^2 \quad \forall \quad \tilde{\mathbf{c}}_{i_{\text{svq}}} \in \tilde{\mathcal{X}}_{\text{svq}} \tag{4.13}$$

$$i_{Q,g} = \underset{0 \le i_g < N_g}{\arg\min} \, |g - \tilde{g}_{i_g}|^2 \quad \forall \quad \tilde{g}_{i_g} \in \tilde{\mathcal{X}}_g. \tag{4.14}$$

Since $\mathbf{c}$ and $\tilde{\mathbf{c}}_{i_{\text{svq}}}$ have unit absolute value, (4.13) can be rewritten as

$$i_{Q,\text{svq}} = \underset{0 \le i_{\text{svq}} < N_{\text{svq}}}{\arg\max} \; \mathbf{c}^T \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}} \tag{4.15}$$

which is often used in the literature.
Both indices $i_{Q,g}$ and $i_{Q,\text{svq}}$ are combined to produce the overall codeword $i_{Q,\text{lsvq}}$ with either

$$i_{Q,\text{lsvq}} = i_{Q,g} \cdot N_{\text{svq}} + i_{Q,\text{svq}} \qquad \text{or} \qquad i_{Q,\text{lsvq}} = i_{Q,\text{svq}} \cdot N_g + i_{Q,g}. \tag{4.16}$$

In practical applications, the overall codeword $i_{Q,\text{lsvq}}$ is transmitted from the encoder to the decoder to reconstruct the quantized vector. For the sake of simplicity, however, the transmission of the codeword is not part of Figure 4.3.
Multiplying the spherical codevectors with different quantization reconstruction levels for the gain factor corresponding to the entries in the codebook $\tilde{\mathcal{X}}_g$ according to (4.9), the overall codebook $\tilde{\mathcal{X}}_{\text{lsvq}}$ is the aggregation of spherical codevectors on spheres with different radius,

$$\tilde{\mathcal{X}}_{\text{lsvq}} := \bigcup_{\tilde{g} \in \tilde{\mathcal{X}}_g} \tilde{g} \cdot \tilde{\mathcal{X}}_{\text{svq}} \tag{4.17}$$

In Figure 4.4 an example distribution of overall codevectors (dots) is shown for $L_v = 2$. In addition, also the quantization cell bounds with respect to the parallel combination of the two quantizers (Figure 4.3) are shown.

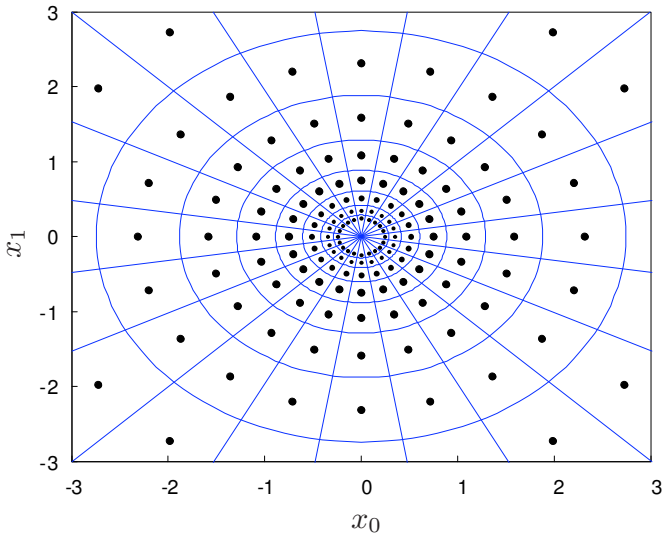


Figure 4.4: Example codevectors and quantization cells for the parallel approach of LSVQ for $L_v = 2$.

4.2.3 A Qualitative Analysis for High Bit Rates

In the first part of the analysis of LSVQ, qualitative results are derived for the assumption of a high bit rate and hence large number of spherical codevectors. Given the spherical codebook $\tilde{\mathcal{X}}_{\text{svq}}$, the *spherical quantization cells* are defined as

$$C_{\tilde{\mathbf{c}}} := \{\mathbf{c} \in \mathcal{S}_{L_v}^{(1.0)} : \|\mathbf{c} - \tilde{\mathbf{c}}\| \leq \|\mathbf{c} - \tilde{\mathbf{c}}'\| \ \forall \ \tilde{\mathbf{c}}, \tilde{\mathbf{c}}' \in \tilde{\mathcal{X}}_{\text{svq}} \text{ and } \tilde{\mathbf{c}}' \neq \tilde{\mathbf{c}}\} \tag{4.18}$$

with $\tilde{\mathbf{c}}$ and $\tilde{\mathbf{c}}'$ representing two independent codevector for which the index i_{svq} is skipped for the sake of simplicity. In analogy to the assumptions for VQ in general in [Ger79] it is assumed that the complete surface is covered by quantization cells of identical shape (the equivalent of the tessellation, restricted to be located on the unit sphere surface). The area content of each spherical quantization cell $S_{C_{\tilde{\mathbf{c}}}}^{(I)}$ can be computed as a function of the overall surface content of a *unit* sphere (4.5), divided by the number of spherical codevectors N_{svq}:

$$B := S_{C_{\tilde{\mathbf{c}}}}^{(I)} = \frac{S_{\mathcal{S}_{L_v}}^{(1.0)}}{N_{\text{svq}}} = V_{\mathcal{S}_{L_v}}^{(1.0)} \cdot \frac{L_v}{N_{\text{svq}}}. \tag{4.19}$$

An example with four spherical quantization cells on the surface of a sphere is shown for $L_v = 3$ on the left side of Figure 4.5.
Considering the SQ for the gain factor g, the quantization intervals are defined as

$$C_{\tilde{g}} := \{g \in \mathbb{R}^+ : |g - \tilde{g}| < |g - \tilde{g}'| \ \ \forall \ \ \tilde{g}, \tilde{g}' \in \tilde{\mathcal{X}}_g \text{ and } \tilde{g}' \neq \tilde{g}\}. \tag{4.20}$$

For the construction of the overall codevectors $\tilde{\mathbf{x}}$ according to (4.17), the spherical codevectors $\tilde{\mathbf{c}}$ are combined with different quantization reconstruction levels $\tilde{g}$, and *overall quantization cells* $C_{\tilde{\mathbf{x}}}$ result, defined as

$$C_{\tilde{\mathbf{x}}} := \{\mathbf{x} \in \mathbb{R}^{L_v} : (\mathbf{c} = \frac{\mathbf{x}}{\|\mathbf{x}\|} \in C_{\tilde{\mathbf{c}}}) \wedge (g = \|\mathbf{x}\| \in C_{\tilde{g}})\}. \tag{4.21}$$

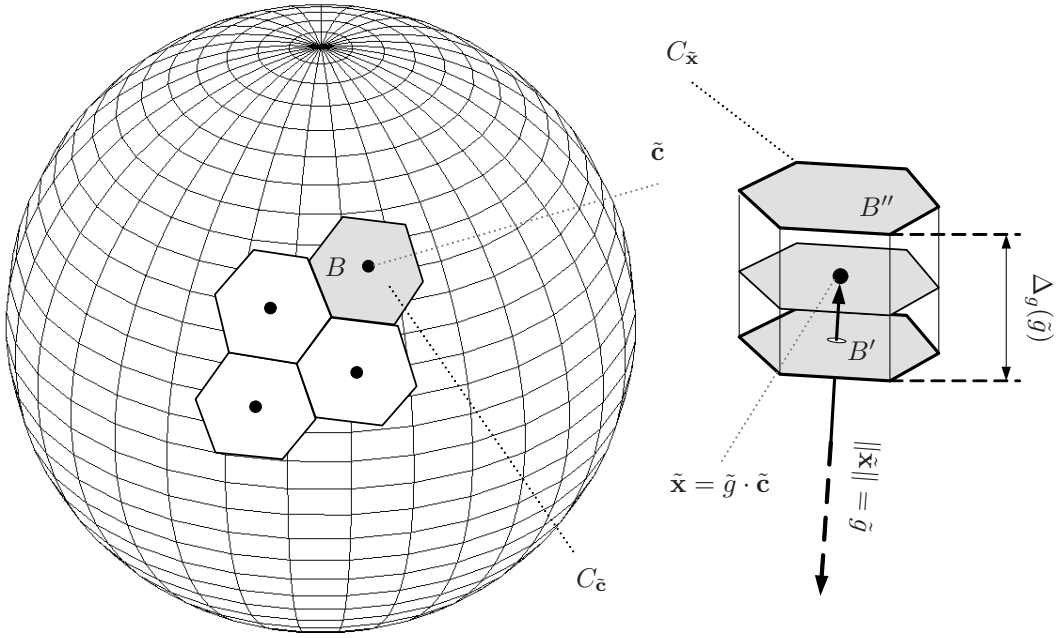


Figure 4.5: Spherical quantization cells located on the surface of a sphere (left side) and example overall quantization cell (right side) for $L_v = 3$.

On the right side of Figure 4.5, an example overall quantization cell is shown qualitatively for $L_v = 3$. Compared to the "flat" spherical quantization cells (denoted as B in the figure and in (4.19) for simplicity), the overall quantization cells also have a height in radial direction which is the distance between adjacent radius quantization interval bounds or, in other words, the size of the radius quantization interval, denoted as $\Delta_g(\tilde{g})$. The bottom and the upper side surface areas of that cell, B' and B'', respectively, are scaled versions of the spherical quantization cells B on the unit sphere surface. For high bit rates a large number of spherical codevectors exists so that the curvature of the sphere can be neglected, and the quantization intervals $\Delta_g(\tilde{g})$ are small. In this case, the bottom and the upper side surface area of the overall quantization cell are approximately identical (therefore $B' \approx B''$ in Figure 4.5). The overall codevector $\tilde{\mathbf{x}}$ is located in the center of the quantization cell and at the same time on the surface of a scaled version of the unit sphere with the radius $\tilde{g} = \|\mathbf{x}\|$. The respective surface area of the scaled unit sphere with the radius $\tilde{g}$ is the unit sphere surface area content multiplied with the factor $\tilde{g}^{(L_v-1)}$ (4.5). As a consequence of this, also the bottom and the upper side surface areas of each cell are $B' \approx B'' \approx B \cdot \tilde{g}^{(L_v-1)}$ (4.19). The complete vector space is filled by quantization cells which are scaled versions of the cell shown in Figure 4.5. The cell volume can be derived as a function of the gain quantization reconstruction level as

$$V_{C_{\tilde{\mathbf{x}}}}(\tilde{g}) \approx \underbrace{B \cdot \tilde{g}^{(L_v-1)}}_{B' \approx B''} \cdot \Delta_g(\tilde{g}) = S^{(I)}_{C_{\tilde{\mathbf{c}}}} \cdot \tilde{g}^{(L_v-1)} \cdot \Delta_g(\tilde{g}). \tag{4.22}$$

Assuming that the gain SQ is operated in the logarithmic area of the A-Law compression curve, $\Delta_g(\tilde{g})$ is approximately the quantization error related to A-Law

quantization of the radius [JN84],

$$\Delta_g(\tilde{g}) \approx \frac{(1+\ln(A))}{N_g} \cdot \tilde{g} \tag{4.23}$$

with A as the A-Law quantization constant. Hence, with (4.23), (4.22), (4.19), and (4.3), the overall quantization cell volume is approximately

$$V_{C_{\tilde{\mathbf{x}}}}(\|\tilde{\mathbf{x}}\|) \approx \frac{2 \cdot \pi^{L_v/2} \cdot (1+\ln(A))}{\Gamma(L_v/2)} \cdot \frac{(\|\tilde{\mathbf{x}}\|)^{L_v}}{N_{\text{svq}} \cdot N_g}, \tag{4.24}$$

expressed as a function of the absolute value of the corresponding overall codevector $\tilde{\mathbf{x}}$ with absolute value $\|\tilde{\mathbf{x}}\| = \tilde{g}$. From this cell volume, the normalized quantizer point density function can be derived as

$$\lambda(\tilde{\mathbf{x}}) \approx \frac{1}{V_{C_{\tilde{\mathbf{x}}}}(\tilde{g}) \cdot N_g \cdot N_{\text{svq}}} = \frac{\Gamma(L_v/2)}{2 \cdot \pi^{L_v/2} \cdot (1+\ln(A))} \cdot (\|\tilde{\mathbf{x}}\|)^{-L_v}. \tag{4.25}$$

Given $p(\mathbf{x})$ as the PDF of the input signal source, it is a common assumption in high bit rate VQ that $p(\mathbf{x}) \approx p(\tilde{\mathbf{x}})$ and that $\lambda(\tilde{\mathbf{x}})$ is a continuous function. According to [Ger79], it follows with (4.12) that the mean of the overall quantization distortion can be computed from (4.25) as

$$D_{\text{lsvq}}^{(I)} = N_{\text{lsvq}}^{-2/L_v} \cdot C_{\text{lsvq}} \cdot \int\limits_{\tilde{\mathbf{x}} \in \mathbb{R}^{L_v}} \frac{p(\tilde{\mathbf{x}})}{(\lambda(\tilde{\mathbf{x}}))^{2/L_v}} \mathrm{d}\tilde{\mathbf{x}}. \tag{4.26}$$

Correspondingly, the SNR can be computed as

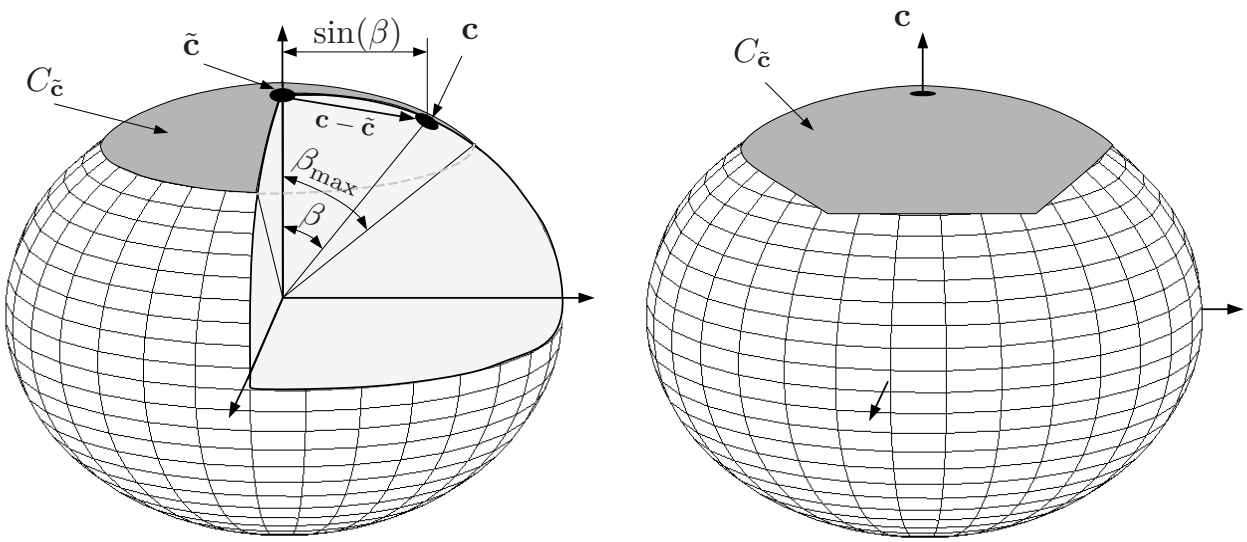
$$\text{SNR}_{\text{lsvq}}^{(I)} = \frac{N_{\text{lsvq}}^{2/L_v}}{C_{\text{lsvq}} \cdot \pi} \cdot \left(\frac{\Gamma(L_v/2)}{2 \cdot (1+\ln(A))} \right)^{2/L_v} \cdot \underbrace{\frac{\int\limits_{\mathbf{x} \in \mathbb{R}^{L_v}} p(\mathbf{x}) \cdot \|\mathbf{x}\|^2 \cdot \mathrm{d}\mathbf{x}}{\int\limits_{\tilde{\mathbf{x}} \in \mathbb{R}^{L_v}} p(\tilde{\mathbf{x}}) \cdot \|\tilde{\mathbf{x}}\|^2 \cdot \mathrm{d}\tilde{\mathbf{x}}}}_{\approx 1} \tag{4.27}$$

with constant C_{lsvq} in (4.26) and (4.27) depending on the (yet unknown) shape of the overall quantization cells. As a conclusion, the SNR related to LSVQ is approximately independent of the PDF of the input signal. [1]

4.2.4 Quantitative Results

Based on the proof of a constant SNR in the previous section (4.27), w.l.o.g. only a single overall quantization cell, a "prototype cell", is considered in the following

[1] Note that for the case of high vector dimensions and a non-logarithmic SQ of the gain (e.g. a LMQ (Section 3.1.1.4) as often used in speech coding), it can be shown that for high bit rates gain-shape VQ leads to a nearly constant SNR. For example, in the denominator of the second part of 4.27 the integral would be a function of $\|\mathbf{x}\|^{2 \cdot (L_v-1)/L_v}$ rather than $\|\mathbf{x}\|^2$ for uniform SQ of the gain).



a) "Spherical cap" quantization cells

b) "Realistic" quantization cell

Figure 4.6: Types of spherical quantization cells, $L_v = 3$.

with the corresponding overall codevector located on the surface of the unit sphere. A quantitative expression for the SNR rather than the qualitative one in (4.27) with the unknown constant C_{lsvq} shall be derived. In the first step, the quantization distortion solely related to Q_{svq} will be computed. This result will then be combined with the distortion related to Q_g in the second step.
The **per vector** distortion D^* (3.32) rather than the **per vector coordinate** distortion D (3.31) and, correspondingly, the per vector rather than the per vector coordinate mean of the squared absolute value of the signal to be quantized will be considered in the following. Since the difference is only a multiplication with the constant L_v, computing the SNR from the per vector measures instead of the per vector coordinate measures does not affect the overall result (in the end, the multiplication cancels out, and this procedure significantly simplifies the notation.).

4.2.4.1 Analysis of the "Idealized" SVQ

Since the exact shape of the spherical quantization cells is unknown, in analogy to the sphere bound for VQ in general [Ger79], we define an "idealized" SVQ to be composed of "spherical cap" quantization cells as illustrated on the left side of Figure 4.7 for $L_v = 3$. A single spherical cap is shown in Figure 4.6 a). The spherical codevector $\tilde{\mathbf{c}}$ is located in the center (at the north pole), and the angular radius $\beta_{\max}$ determines the size of the spherical cap. The spherical cap area content [Ham96] is

$$S_{C_{\tilde{\mathbf{c}}}}^{(II)} = V_{\mathcal{S}_{L_v-1}}^{(1.0)} \cdot (L_v{-}1) \cdot \int_0^{\beta_{\max}} (\sin(\beta))^{(L_v-2)} \mathrm{d}\beta. \tag{4.28}$$

Note that the upper index (II) is used here to point out the difference to the cell area content in (4.19).

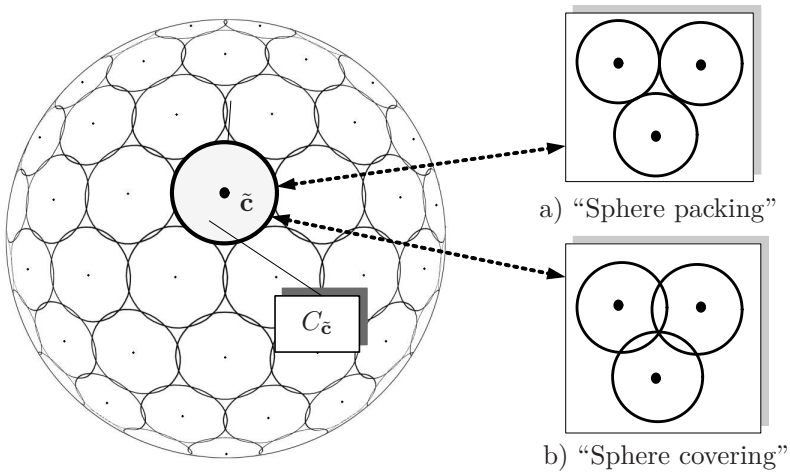


Figure 4.7: Covering of unit sphere surface by "spherical cap" quantization cells.

Given a normalized vector $\mathbf{c}$ which is located inside the spherical cap area as shown in the example of Figure 4.6, the absolute value of the quantization error can be computed by means of the cosine rule as a function of β,

$$\|\mathbf{c} - \tilde{\mathbf{c}}\| = 2 \cdot \sin(\beta/2). \tag{4.29}$$

Assuming that the number of spherical cap cells is high enough so that the distribution of $\mathbf{c}$ within the spherical cap quantization cell is uniform (constant PDF $p(\mathbf{c})$) and with (4.28), the per vector distortion as a function of angular radius $\beta_{\max}$ is

$$D^{*(II)}_{\text{svq}} = \int_{C_{\tilde{\mathbf{c}}}} p(\mathbf{c}) \cdot \|\mathbf{c} - \tilde{\mathbf{c}}\|^2 \mathrm{d}\mathbf{c} = \frac{\int_0^{\beta_{\max}} (2 \cdot \sin(\frac{\beta}{2}))^2 \cdot (\sin(\beta))^{(L_v-2)} \mathrm{d}\beta}{\int_0^{\beta_{\max}} (\sin(\beta))^{(L_v-2)} \mathrm{d}\beta} \tag{4.30}$$

with

$$p(\mathbf{c}) = \text{const} = 1/S^{(II)}_{C_{\tilde{\mathbf{c}}}} \qquad \text{so that} \int_{\mathbf{c} \in C_{\tilde{\mathbf{c}}}} p(\mathbf{c}) \mathrm{d}\mathbf{c} = 1. \tag{4.31}$$

$\beta_{\max}$ is the unknown parameter in (4.30) so far. In order to compute $\beta_{\max}$, it is assumed that the complete unit sphere surface is covered by spherical cap quantization cells as shown on the left side of Figure 4.7. Given the angular radius such that they do not overlap, the spherical caps do only cover a fraction of the complete surface. This approach is the analogon to "sphere packing" in the literature [CS93] and is demonstrated by Figure 4.7 a). Given a large number of codevectors, the sphere surface can be assumed to be covered by approximately flat $(L_v - 1)$-dimensional quantization cells. It is shown in [Ham96] that in this case the density $\delta \leq 1$ from [CS93] for dimension $(L_v - 1)$ is approximately the ratio between the area covered by spherical caps and the unit sphere surface area.
If no uncovered space is allowed, the spherical caps overlap as shown in Figure

4.7 b) which is the analogon to "sphere covering" in the literature. In analogy to the definition of the density, the thickness $\theta \geq 1$ is defined as the proportion of the overall space covered by the spherical caps in relation to the surface area of the unit sphere.
It was described in Section 3.2.1.1 that spheres as the optimal shape of quantization cells lead to the lowest normalized inertia (3.40) which is the basis for the derivation of the "sphere lower bound" for VQ in general. In SVQ, the quantization cells are restricted to be located on the surface of the unit sphere. In analogy to the *sphere* as the optimum cell shape for VQ, the *spherical cap* is therefore identified as the optimal cell shape for SVQ.
If the angular radius $\beta_{\max}$ is determined according to the "sphere packing" assumption, the distortion (4.30) is certainly lower than the distortion achievable by any distribution of codevectors on the sphere surface, denoted as the "realistic" SVQ: The spherical quantization cells of the "idealized" SVQ have the optimal shape and also are smaller than those achievable by any "realistic" SVQ. In contrast to this, in the "sphere covering" approach, even though the cell shape is optimal in the sense to minimize the normalized inertia, the distortion (4.30) is not a bound because the cells are larger than those of a "realistic" SVQ.
The best option to compute $\beta_{\max}$, however, is to define that the spherical caps do not overlap and cover the complete surface at the same time. This assumption is obviously unrealistic as shown for the example $L_v = 3$ in Figure 4.7 but is the basis for a more accurate bound than that related to the "sphere packing" assumption: Since the area content of the spherical quantization cells of the best "realistic" SVQ may be identical to those of the "idealized" SVQ, the "idealized" SVQ has no benefit due to smaller spherical quantization cells. Nevertheless, the distortion achievable by the "idealized" SVQ is lower since it benefits from the optimal spherical quantization cell shape which can not be achieved by a "realistic" SVQ for finite dimensions. An example illustrates this on the right side of Figure 4.6 for $L_v = 3$ where a "realistic" spherical quantization cell is shown which has the same cell content as the "idealized" spherical cap on the left side but a suboptimal cell shape. By computing the angle $\beta_{\max}$ according to this assumption, equation (4.30) turns out to be a lower bound which is more realistic than the lower bound for the "sphere packing" approach. Taking this into account, $\beta_{\max}$ can be computed from

$$S^{(1.0)}_{\mathcal{S}_{Lv}} = N_{\text{svq}} \cdot S^{(II)}_{C_{\tilde{\mathbf{c}}}}. \tag{4.32}$$

Substituting (4.28) in (4.32) yields an equation from which the angular radius $\beta_{\max}$ can be determined. Due to the integral, a direct computation is not straight forward. It can, however, be solved numerically, e.g., by means of *Newtons Method*. Given the computed angular radius $\beta_{\max}$, the quantization distortion can be calculated by numerically solving the integrals in (4.30). From the distortion, the SNR related

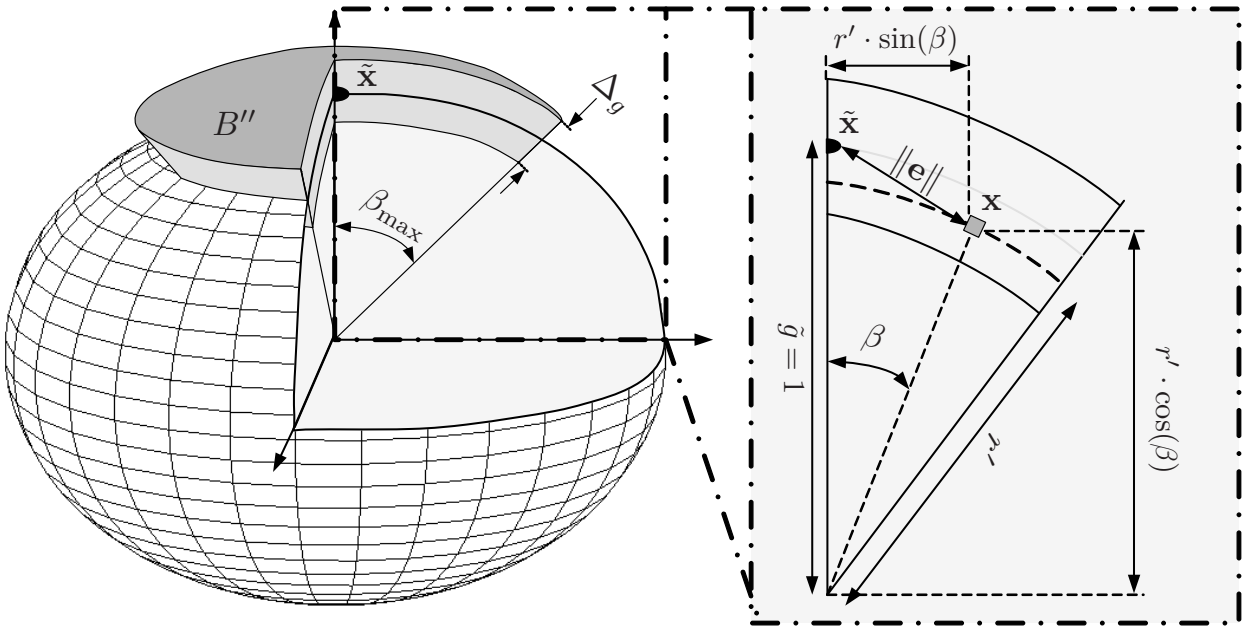


Figure 4.8: L_v-dimensional quantization cell and quantization error vector $\mathbf{e}$.

solely to Q_{svq} is

$$\text{SNR}_{\text{svq}}^{(II)} = \frac{E\{\|\mathbf{c}\|^2\}}{D_{\text{svq}}^{*(II)}} = \frac{\int_0^{\beta_{\max}} (\sin(\beta))^{(L_v-2)} \cdot \mathrm{d}\beta}{\int_0^{\beta_{\max}} (2 \cdot \sin(\beta/2))^2 \cdot (\sin(\beta))^{(L_v-2)} \cdot \mathrm{d}\beta} \tag{4.33}$$

since the mean of the squared absolute values of the normalized signal vectors $\mathbf{c}$ is

$$E\{\|\mathbf{c}\|^2\} = 1. \tag{4.34}$$

4.2.4.2 Analysis of the "Idealized" LSVQ

As explained earlier, the combination of Q_{svq} and Q_g leads to overall quantization cells with a bottom and an upper side area and a height. Compared to Figure 4.5, an example prototype overall quantization cell of the "idealized" LSVQ is shown for $L_v = 3$ in Figure 4.8 where also the curvature of the sphere is considered. In order to exactly compute the content of the bottom and the upper side content (B'' in the figure), the area content from (4.28) must be generalized for an arbitrary sphere radius r'

$$S_{C_{\tilde{\mathbf{c}}}}^{(II)}(r') = V_{\mathcal{S}_{L_v-1}}^{(1.0)} \cdot (L_v - 1) \cdot \int_0^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \tag{4.35}$$

It is assumed that the quantization reconstruction value for the gain factor related to the overall codevector $\tilde{\mathbf{x}}$, $\tilde{g} = 1.0$, is in the middle of the quantization interval $\Delta_g := \Delta_g(\tilde{g} = 1.0)$ so that the sphere radius is in the range of

$$1 - \frac{\Delta_g}{2} \leq r' \leq 1 + \frac{\Delta_g}{2}. \tag{4.36}$$

Correspondingly, the volume of the overall "prototype" quantization cell is

$$\begin{aligned} V_{C_{\tilde{\mathbf{x}}}} &= \int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} S_{C_{\tilde{\mathbf{c}}}}(r')\mathrm{d}r' \\ &= V^{(1.0)}_{\mathcal{S}_{L_v-1}} \cdot (L_v-1) \cdot \int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int_{0}^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'. \end{aligned} \tag{4.37}$$

With respect to all vectors $\mathbf{x}$ inside this cell, the distortion is

$$\begin{aligned} D^{*(II)}_{\mathrm{lsvq}} &= E\{\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\} \\ &= \int_{\mathbf{x}\in C_{\tilde{\mathbf{x}}}} p(\mathbf{x}) \cdot \|\mathbf{x}-\tilde{\mathbf{x}}\|^2 \mathrm{d}\mathbf{x} = \int_{\mathbf{x}\in C_{\tilde{\mathbf{x}}}} p(\mathbf{x}) \cdot \|\mathbf{e}\|^2 \mathrm{d}\mathbf{x} \end{aligned} \tag{4.38}$$

with the error vector $\mathbf{e} = \mathbf{x} - \tilde{\mathbf{x}}$ as illustrated by the cut-out on the right side of Figure 4.8. Further, it is assumed that the number of overall codevectors is high enough so that the PDF $p(\mathbf{x})$ of vectors $\mathbf{x}$ within the overall quantization cell is constant (analog to (4.31)):

$$p(\mathbf{x}) = \mathrm{const} = \frac{1}{V_{C_{\tilde{\mathbf{x}}}}}. \tag{4.39}$$

With respect to Figure 4.8 or by applying the cosine rule, the absolute value of the error related to all vectors $\mathbf{x}$ which are located inside the overall quantization cell, $\|\mathbf{e}\|^2$, can be expressed as a function of β and r':

$$\|\mathbf{e}\|^2 = (r' \cdot \sin(\beta))^2 + (1 - r' \cdot \cos(\beta))^2 \tag{4.40}$$

Substituting this in (4.38) yields the overall distortion,

$$D^{*(II)}_{\mathrm{lsvq}} = \frac{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int_{0}^{\beta_{\max}} ((r' \cdot \sin(\beta))^2 + (1 - r' \cdot \cos(\beta))^2)(r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'}{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int_{0}^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'} \tag{4.41}$$

For the computation of the SNR, the variance of the signal $\mathbf{x}$ with respect to all vectors located in the overall quantization cell is

$$\begin{aligned} E\{\|\mathbf{x}\|^2\} &= \int\limits_{\mathbf{x}\in V_{C_{\tilde{\mathbf{x}}}}} p(\mathbf{x})\|\mathbf{x}\|^2 \mathrm{d}\mathbf{x} \\ &= \frac{\int\limits_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int\limits_{0}^{\beta_{\max}} r'^2 \cdot (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'}{\int\limits_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int\limits_{0}^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'} \end{aligned} \tag{4.42}$$

with $p(\mathbf{x})$ from (4.39). Combining(4.41) and (4.42) leads to the SNR

$$\mathrm{SNR}_{\mathrm{lsvq}}^{(II)} = \frac{\int\limits_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int\limits_{0}^{\beta_{\max}} r'^{L_v+1} \cdot (\sin(\beta))^{(L_v-2)} \cdot \mathrm{d}\beta \cdot \mathrm{d}r'}{\int\limits_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int\limits_{0}^{\beta_{\max}} ((r' \cdot \sin(\beta))^2 + (1 - r' \cdot \cos(\beta))^2)(r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'}. \tag{4.43}$$

The final result is based on computations for the "prototype" overall quantization cell. Because it was assumed that all spherical cells located around the unit sphere surface are identical and due to (4.27), this result can be generalized for all cells and hence LSVQ in general.
For a numerical evaluation of the presented results, given the number of spherical codevectors N_{svq}, the angular radius $\beta_{\max}$ can be calculated from (4.28) and (4.32) in the first step. In the second step, with (4.23) and given the number of quantization reconstruction levels N_g, the SNR related to LSVQ can be calculated by numerically solving the integrals in (4.43). The derived expressions for the SNR related to SVQ (4.33) and to LSVQ (4.43) are upper SNR bounds if the integrals can be solved with sufficient precision. Note, however, that the assumption was made for the sake of simplicity that the quantization reconstruction level $\tilde{g}$ is in the middle of the gain quantization intervals. Due to the curvature of the sphere, the respective codevector $\tilde{\mathbf{x}}$ is not in the center of gravity of the cell which is suboptimal.

4.2.4.3 High Rate Approximations

In the following, approximations are introduced for high bit rate assumptions:

$$r' \approx 1.0 \tag{4.44}$$

$$r' \cdot \sin(\beta) \approx \sin(\beta) \approx \beta \tag{4.45}$$

$$1 - r' \cdot \cos(\beta) \approx 1 - r'. \tag{4.46}$$

These simplifications are related to the negligence of the curvature of the sphere and enable to transform (4.35) and (4.32) into an expression to compute $\beta_{\max}$ analytically as a function of N_{svq}[2],

$$\beta_{\max} = \left(\frac{2\sqrt{\pi} \cdot \Gamma(\frac{L_v+1}{2})}{\Gamma(\frac{L_v}{2}) \cdot N_{\text{svq}}} \right)^{\frac{1}{L_v-1}}. \tag{4.47}$$

The overall per vector distortion (4.41) for high bit rates is approximately[3]

$$\begin{aligned} D^{*(III)}_{\text{lsvq}} = E\{\|\mathbf{x}-\tilde{\mathbf{x}}\|^2\} &\approx \frac{\int_0^{\beta_{\max}} \beta^{L_v} \mathrm{d}\beta}{\int_0^{\beta_{\max}} \beta^{(L_v-2)} \mathrm{d}\beta} + \frac{\Delta_g^2}{12} \\ &= \frac{L_v-1}{L_v+1} \cdot \beta_{\max}^2 + \frac{\Delta_g^2}{12} \\ &= E\{\|\mathbf{c}-\tilde{\mathbf{c}}\|^2\} + E\{|g-\tilde{g}|^2\}. \end{aligned} \tag{4.48}$$

The quantization errors related to the SVQ and the gain SQ part of LSVQ are hence independent. This result is very intuitive since the error vector related to the spherical codevectors (in tangential direction) and the quantization of the radius (in radial direction) are orthogonal if the curvature of the sphere is negligible.
Substituting $\beta_{\max}$ from (4.47) and with the distortion related to A-Law SQ [JN84],

$$D_g = \frac{C_g}{N_g^2}, \tag{4.49}$$

(4.48) can be written as

$$D^{*(III)}_{\text{lsvq}} = \underbrace{C_{\text{svq}} \cdot N_{\text{svq}}^{-\frac{2}{L_v-1}}}_{D^{*(III)}_{\text{svq}}} + \underbrace{C_g \cdot N_g^{-2}}_{Dg} \tag{4.50}$$

with the constants

$$C_{\text{svq}} = \frac{L_v-1}{L_v+1} \cdot \left(\frac{2\sqrt{\pi} \cdot \Gamma(\frac{L_v+1}{2})}{\Gamma(\frac{L_v}{2})} \right)^{\frac{2}{L_v-1}} \quad \text{and} \quad C_g = \frac{(1+\ln(A))^2}{12} \tag{4.51}$$

By using the approximations rather than the exact solution, with respect to (4.41)[4], it can not be guaranteed that (4.50) is a bound. Therefore, all high rate results must be considered as a performance estimate for LSVQ rather than a bound.

[2]The derivation of this equation is in detail explained in Appendix A.1

[3]The derivation of this equation is in detail explained in Appendix A.2

[4]By introducing the approximation for the sine function, the numerator and the denominator are diminished simultaneously so that it is unknown whether the overall expression (4.41) is increased or decreased.

4.2.4.4 Optimal Bit Allocation

In order to find the optimal distribution of the overall bit rate to Q_g and Q_{svq}, the overall distortion (4.50) is minimized in a Lagrangian optimization, given the constraint $N_{\text{lsvq}} = N_{\text{svq}} \cdot N_g$ in (4.12). The auxiliary function is

$$\chi = \frac{C_{\text{svq}}}{N_{\text{svq}}^{\frac{2}{L_v-1}}} + \frac{C_g}{N_g^2} + \lambda \cdot (N_g \cdot N_{\text{svq}} - N_{\text{lsvq}}). \tag{4.52}$$

Setting its partial derivatives with respect to N_g and N_{svq} to zero and with (4.50) an intermediate result[5] is derived as

$$D_g \stackrel{!}{=} \frac{D_{\text{svq}}^{*(III)}}{L_v - 1}. \tag{4.53}$$

The optimal number of spherical codevectors and gain quantization reconstruction levels is finally computed as

$$N_{\text{svq}} = \left(\frac{1}{L_v - 1} \cdot \frac{C_{\text{svq}}}{C_g} \right)^{\frac{L_v-1}{2 \cdot L_v}} \cdot N_{\text{lsvq}}^{\frac{L_v-1}{L_v}} \tag{4.54}$$

$$N_g = \left((L_v - 1) \cdot \frac{C_g}{C_{\text{svq}}} \right)^{\frac{L_v-1}{2 \cdot L_v}} \cdot N_{\text{lsvq}}^{\frac{1}{L_v}}. \tag{4.55}$$

As the considered "prototype" quantization cell is located around the surface of the unit sphere, for high bit rates, the per vector variance of signal vector $\mathbf{x}$ is approximately $E\{\|\mathbf{x}\|^2\} \approx E\{\|\mathbf{c}\|^2\} = 1$ since $\tilde{g} = 1$. After substituting (4.54) and (4.55) in (4.50), the overall logarithmic SNR in dB as a function of the overall bit rate per vector coordinate $R_{\text{eff, lsvq}}$ for high bit rates is[6]

$$\begin{aligned} \text{SNR}_{\text{lsvq}}^{(III)}|_{\text{dB}} &= 6.02 \cdot R_{\text{eff, lsvq}} \\ &\quad - 10 \log_{10} \left(\frac{L_v}{(L_v - 1)^{\frac{L_v-1}{L_v}}} \cdot \left[2\sqrt{\pi} \frac{\Gamma(\frac{L_v-1}{2})}{\Gamma(\frac{L_v}{2})} \right]^{\frac{2}{L_v}} \left[\frac{(1 + \ln(A))^2}{12} \right]^{\frac{1}{L_v}} \right). \end{aligned} \tag{4.56}$$

Considering the asymptotic case for infinite dimensions, it can be shown that

$$\lim_{L_v \to \infty} \text{SNR}_{\text{lsvq}}^{(III)}|_{\text{dB}} = 6.02 \cdot R_{\text{eff,lsvq}}. \tag{4.57}$$

As a conclusion, LSVQ asymptotically reaches the rate distortion function for uncorrelated Gaussian sources for high bit rates and infinite dimensions. Note that an intuitive explanation for this behavior is presented in Section 4.3.6.

[5] The derivation of these equations is in detail explained in Appendix A.3

[6] The derivation of this equation is in detail explained in Appendix A.4

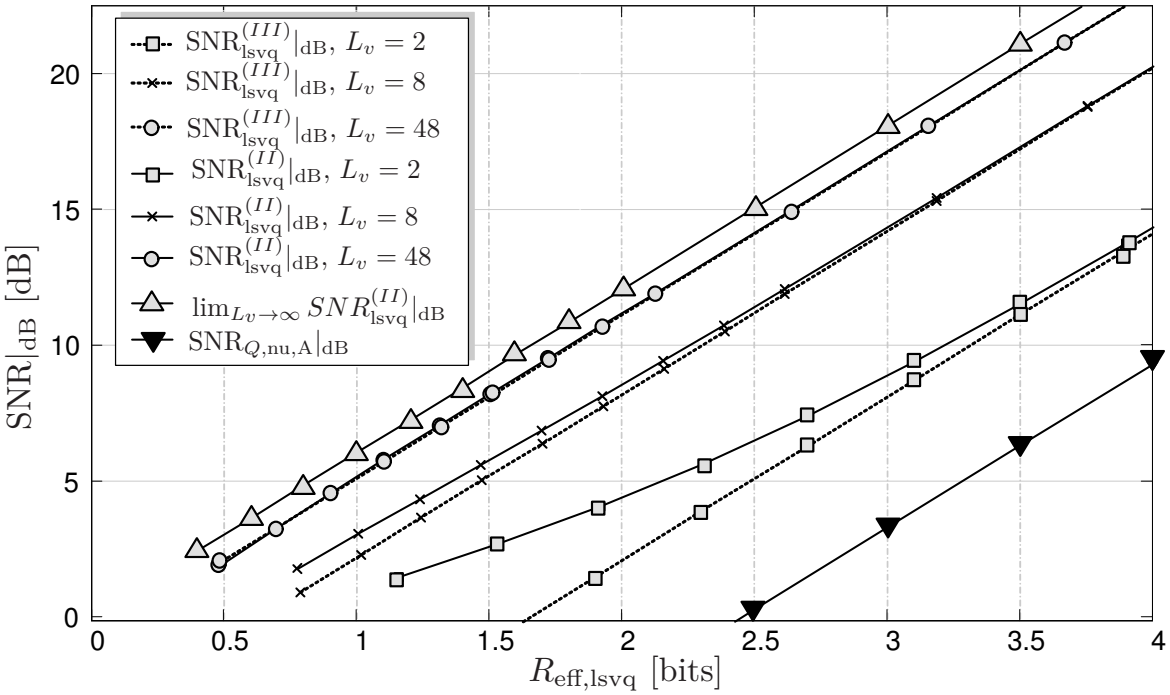


Figure 4.9: LSVQ SNR plots for dimensions $L_v = 2, 8, 48$ over the effective bit rate. In addition, $\mathrm{SNR}_{Q,\mathrm{nu,A}}|_{\mathrm{dB}}$ (3.22) (A-Law non-uniform SQ, LSVQ for $L_v = 1$) and $\lim_{Lv\to\infty} \mathrm{SNR}^{(III)}_{\mathrm{lsvq}}|_{\mathrm{dB}}$ (identical to rate distortion function for memoryless Gaussian sources) are shown as reference curves. All results have been produced for $A = 5000$.

4.3 Evaluation of the Theory

The theoretical results related to SVQ and LSVQ presented previously, denoted as the operational SVQ and LSVQ rate distortion functions, shall be evaluated in the following. For this purpose, SNR plots for different dimensions L_v and effective bit rates $R_{\mathrm{eff,lsvq}}$ and $R_{\mathrm{eff,svq}}$ are presented. In all evaluations the constant A (also refer to Chapter 3.1.1.5) is set to a fixed value of

$$A = 5000. \tag{4.58}$$

This value is chosen as an example here since it has been identified in informal listening tests in the context of the audio codecs described in Chapter 6 as a reasonable trade-off to achieve a high quantization performance as well as that the quantization noise is inaudible in signal pauses.

4.3.1 SNR Plots related to LSVQ

At first SNR plots for the high rate approximations, $\mathrm{SNR}^{(III)}_{\mathrm{lsvq}}|_{\mathrm{dB}}$ according to (4.56), and for the exact result $\mathrm{SNR}^{(II)}_{\mathrm{lsvq}}|_{\mathrm{dB}}$ according to (4.43) are shown for different effective bit rates $R_{\mathrm{eff,svq}}$ and dimensions $L_v = 2$, $L_v = 8$ and $L_v = 48$ in Figure 4.9. For the computation of the exact SNR according to (4.43), in the first step, given a value for N_{svq}, the angular radius $\beta_{\max}$ from (4.28) and (4.32), and the distortion

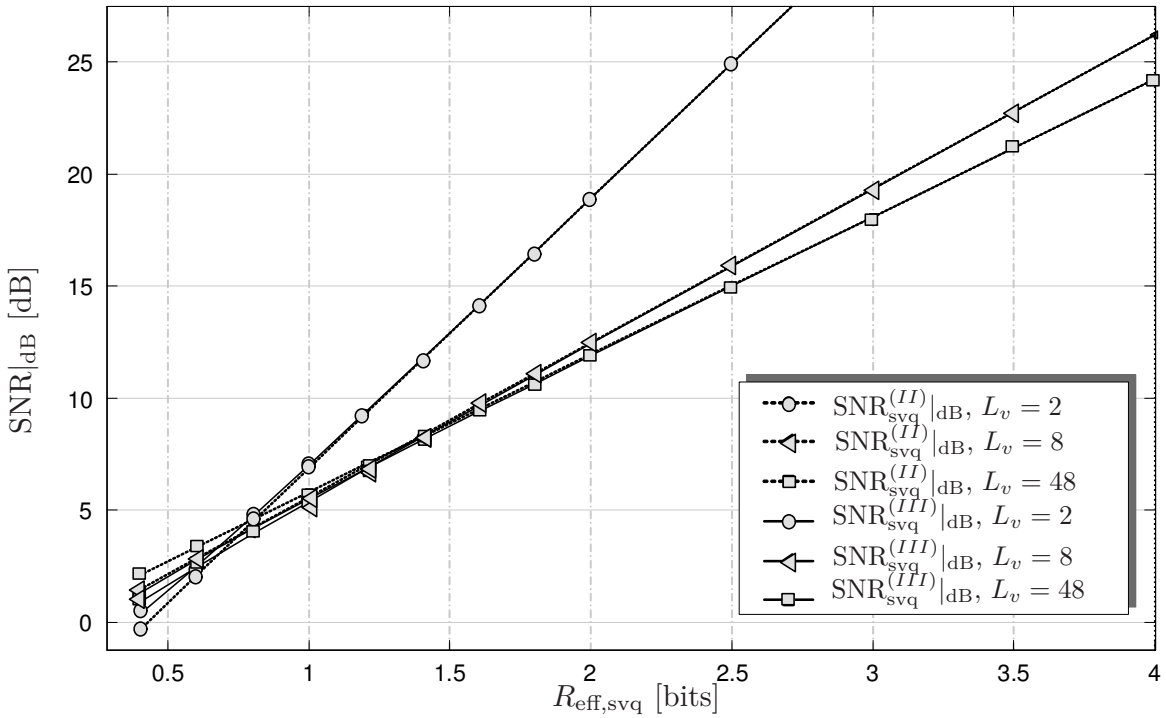


Figure 4.10: SNR for SVQ in dB for dimensions $L_v = 2, 8, 48$ over the effective bit rate $R_{\text{eff,svq}}$.

$D^{*(II)}_{\text{svq}}$ from (4.30) are computed. Based on the intermediate result (4.53) from the derivation of the optimal bit allocation, in the second step, D_g is computed from $D^{*(II)}_{\text{svq}}$ to determine Δ_g in (4.43) and N_g (4.49). The asymptotic value for $L_v \to \infty$ which is also the rate distortion function for memoryless Gaussian sources and the SNR related to A-Law SQ as the special case for LSVQ with $L_v = 1$ are shown as reference curves.
The quantization performance increases for higher dimensions. Compared to the case of $L_v = 1$, already very low values for L_v provide a significant benefit. A clear difference between the high rate approximation and the exact solution is visible for low bit rates.

4.3.2 SNR Plots related to SVQ

A relevant tool for the assessment of the suitability of spherical codes for SVQ is the SNR curve for Q_{svq}. The SNR plots in Figure 4.10 for dimensions $L_v = 2$, $L_v = 8$ and $L_v = 48$ and different effective SVQ bit rates $R_{\text{eff,svq}} = \log_2(N_{\text{svq}})/L_v$ are based on the distortion $D^{*(II)}_{\text{svq}}$ determined according to the exact solution (4.30) with the angular radius computed from (4.28), and the high rate approximation $D^{*(III)}_{\text{svq}}$ as the first part of (4.48). From the distortion, the SNR is computed with (4.34). A difference between the exact SNR and the high rate approximation is only visible for low bit rates. An interesting outcome of the diagram is that the SNR grows faster than the 6-dB-per bit rule for lower values of L_v. The reason for this behavior is that the bit rate per vector coordinate is computed from the

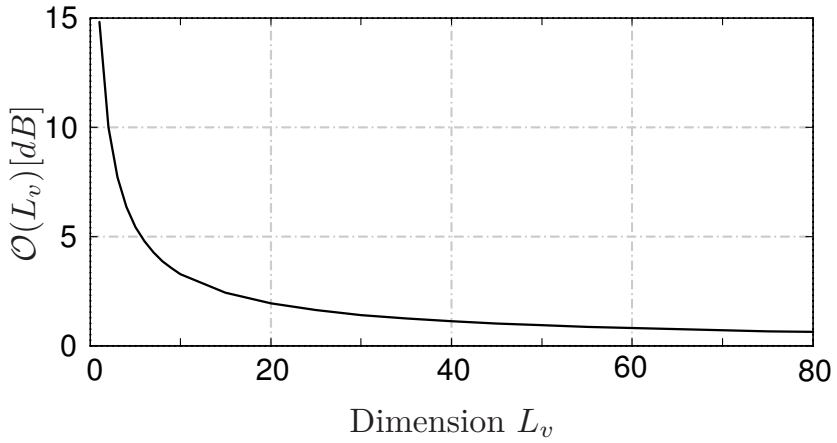


Figure 4.11: Offset of quantization performance of LSVQ in comparison to the 6-dB-per-bit bound.

dimension L_v whereas the SVQ is designed for a $(L_v - 1)$-dimensional surface area and the relative difference between L_v and $(L_v - 1)$ is more significant for lower values of L_v than for higher. For infinite dimension, the SNR would exactly follow the 6-dB-per-bit rule.

4.3.3 SNR over Dimension

In Figure 4.11, the difference of $\mathrm{SNR}_{\mathrm{lsvq}}^{(III)}$ for finite dimensions (4.56) and the asymptotic SNR for infinite dimensions (4.57),

$$\mathcal{O}(L_v) = \lim_{L_v \to \infty} \mathrm{SNR}_{\mathrm{lsvq}}^{(III)}|_{\mathrm{dB}} - \mathrm{SNR}_{\mathrm{lsvq}}^{(III)}|_{\mathrm{dB}}, \tag{4.59}$$

($\mathrm{SNR}_{\mathrm{lsvq}}^{(III)}|_{\mathrm{dB}}$ is a function of L_v) is shown as a function of the vector dimension L_v. For quantization in practice, this result indicates that, on the one side, an increase of performance can be achieved by increasing the dimension L_v. However, on the other side, the highest performance gain due to an increase of the dimension is achieved for low values. Taking into account also that vector quantizers with higher dimensions in general are connected to nearest neighbor codevector search procedures with higher computational complexity and memory consumption, at moderate dimensions already the additional computational effort involved in the increase of the VQ dimension may no longer be justified. Given the results from Section 3.1.2, the asymptotic performance of entropy constrained SQ for high bit rates is approximately outperformed for $L_v > 26$. A variant of LSVQ where large vector dimensions are achieved by means of concatenation of spherical codes to increase the overall SNR is in detail discussed in Section C.3 of the *supplement document* [Krü09].

4.3.4 Visualization of the Spherical Code Quality

A valuable tool for the assessment of the quality of different spherical codes for SVQ is the visualization of the distribution of the quantization error related to Q_{svq} as

it contains information about the average spherical quantization cell shape. Since a direct visualization of multivariate distributions is not possible for $L_v > 2$, the tool is based on indirect measures related to distribution of the *absolute value of the quantization error*. Given a specific SVQ proposal, the histogram of measured absolute values of quantization error vectors in comparison to the theoretical distribution computed for the assumption of spherical cap quantization cells yields a qualitative evaluation criterion.
In order to get the reference of the theoretically achievable distribution of the absolute value of the quantization error for the "idealized" SVQ, it is assumed that the normalized vectors $\mathbf{c}$ are uniformly distributed within each spherical quantization cell with the PDF

$$p(\mathbf{c}) = \begin{cases} \frac{1}{S_{C_{\tilde{\mathbf{c}}}}^{(II)}} & \text{if } \mathbf{c} \in C_{\tilde{\mathbf{c}}} \\ 0 & \text{otherwise} \end{cases} \tag{4.60}$$

with $S_{C_{\tilde{\mathbf{c}}}}^{(II)}$ from (4.28). The distribution of the absolute value of the quantization error vectors can be computed as the probability that a vector $\mathbf{c}$ is located on the surface of a $((L_v-1)$-dimensional) shell around the codevector $\tilde{\mathbf{c}}$ with radius $\|\mathbf{c}-\tilde{\mathbf{c}}\|$ (and of course on the surface of the unit sphere simultaneously)[7]. The $((L_v-1)$-dimensional) shell has a radius of $r_\beta = \sin(\beta)$ so that the theoretical distribution of the absolute value of the quantization error vectors can be written as a function of the angular radius β (see Figure 4.6) as

$$p(\|\mathbf{c}-\tilde{\mathbf{c}}\|) = \begin{cases} p(\mathbf{c}) \cdot S_{\mathcal{S}_{L_v}}^{(r_\beta)} = \frac{(\sin(\beta))^{(L_v-2)}}{\int_0^{\beta_{\max}} (\sin(\beta'))^{(L_v-2)} \mathrm{d}\beta'} & \text{for } 0 \le \beta \le \beta_{\max} \\ 0 & \text{else} \end{cases} . \tag{4.61}$$

An example PDF of the absolute value of the quantization error vector for $L_v = 8$ and $N_{\mathrm{svq}} = 502$ is given in Figure 4.12. Under the idealized assumption that all spherical cap quantization cells are identical, non-overlapping and cover the complete surface of the unit sphere, the PDF increases until the maximum value is reached for $\beta = \beta_{\max}$. This maximum is at the same time the outer boundary of the spherical cap cell. Outside the spherical cap this PDF is of course zero. Given the PDF of the absolute value of the error vector, in analogy to (4.30), the SVQ quantization distortion is

$$E\{\|\mathbf{c}-\tilde{\mathbf{c}}\|^2\} = \int_0^{\beta_{\max}} p(\|\mathbf{c}-\tilde{\mathbf{c}}\|) \cdot \underbrace{(2 \cdot \sin(\beta/2))^2}_{\|\mathbf{c}-\tilde{\mathbf{c}}\|^2} \mathrm{d}\beta \tag{4.62}$$

[7] In a three-dimensional configuration this $((L_v-1)$-dimensional) shell would be circles which concentrically surround the codevectors of each cell and are restricted to be located on the surface of the three-dimensional unit sphere.

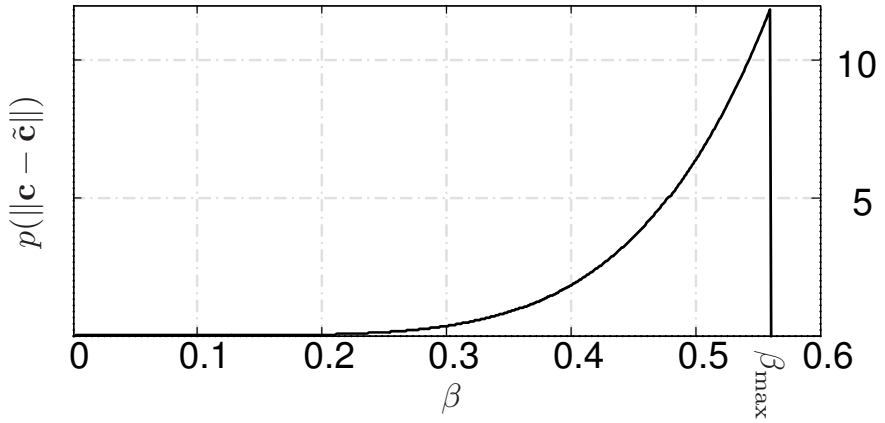


Figure 4.12: Example of the PDF of the absolute value of the quantization error vector $p(\|\mathbf{c}-\tilde{\mathbf{c}}\|)$ over angle β for $L_v = 8$, $A = 5000$, $N_{\text{svq}} = 502$ and $\beta_{\max} = 0.56$.

Given a measured histogram of the absolute value of the quantization error, the assessment of the code quality is based on a comparison to the theoretically achievable distribution given in Figure 4.12. The more similar the measured distribution is to the theoretic curve, the higher is quality of the spherical code for quantization.

4.3.5 Optimal Bit Allocation

The optimal bit allocation from (4.56) is illustrated by Figure 4.13 where *different bit allocation mismatch* situations are investigated for $L_v = 8$ and $A = 5000$. Given a fixed value of N_{svq}, the optimal number of quantization reconstruction levels

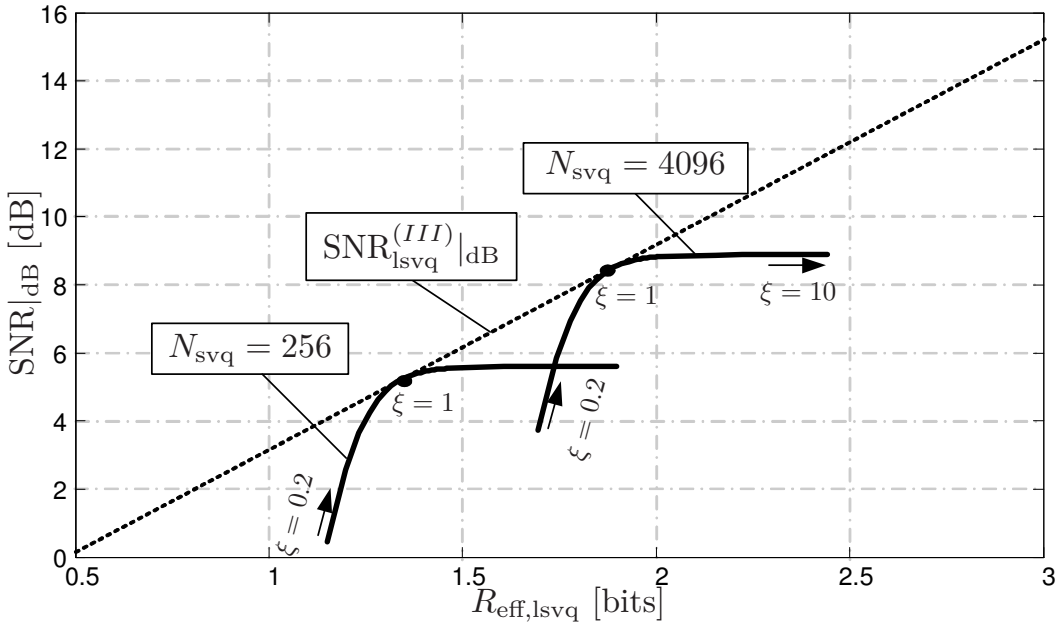


Figure 4.13: SNR for LSVQ with a bit allocation mismatch for $L_v = 8$ and $A = 5000$. The mismatch factor $0.2 \leq \xi \leq 10$ is used to parameterize different points of operation.

for the gain factor can be computed according to (4.54) and (4.55) as N_g. The mismatch is achieved by computing a suboptimal number

$$N'_g = \xi \cdot N_g \tag{4.63}$$

to be used for LSVQ for different values of the mismatch coefficient ξ. Based on N'_g and N_{svq} a bit rate as

$$R_{\text{eff, lsvq}} = 1/L_v \cdot \log_2(N'_g \cdot N_{\text{svq}}) \tag{4.64}$$

and the SNR from the distortion according to (4.50) are computed to produce points of the solid lines in the figure.
Two SNR curves have been determined for $N_{\text{svq}} = 256$ and $N_{\text{svq}} = 4096$ in Figure 4.13, parameterized by the mismatch factor $0.2 \leq \xi \leq 10$ as indicated by the black arrows. In addition the maximum achievable SNR in dB for LSVQ according to (4.56) (which was computed for the assumption of the optimal bit allocation and arbitrary values of N_{svq}) is shown for $L_v = 8$ and different effective bit rates as the dotted line. Obviously, the higher the number N_g, the higher is also the quantization SNR and the overall bit rate. The most *efficient* and therefore optimal configuration in rate distortion sense, however, is achieved at the positions where the two curves touch the dotted line. These points are reached for $\xi = 1.0$.

4.3.6 Plausibility for Infinite Dimension

According to (4.57), for infinite dimension, LSVQ reaches the rate distortion function for Gaussian sources. This result is at the first glance not necessarily obvious since LSVQ was introduced only as an approximation of the optimal codevector density in (4.1) but can be well understood due to the "Sphere Hardening" effect [GS88]: Recalling the results from high rate VQ theory, for infinite dimension, the quantizer codevector density function is

$$\lim_{L_v \to \infty} \lambda_{L_v}(\mathbf{x}) = p(\mathbf{x}) \tag{4.65}$$

This density is related to a vector $\mathbf{x}$ but can be transformed also into a *sphere density* which is a function of the absolute value of vector $\mathbf{x}$ in analogy to Section 4.3.4

$$\lambda_{L_v,\text{sp}}(\|\mathbf{x}\|) = \lambda_{L_v}(\mathbf{x}) \cdot S_{\mathcal{S}_{L_v}}(\|\mathbf{x}\|) \tag{4.66}$$

with $S_{\mathcal{S}_{L_v}}(\|\mathbf{x}\|)$ as the surface area content of a sphere with radius $\|\mathbf{x}\|$. The sphere density $\lambda_{L_v,\text{sp}}(\|\mathbf{x}\|)$ for different dimensions L_v is depicted in Figure 4.14 for the assumption of a Gaussian source with

$$\sigma_x^2 = \frac{1}{L_v} \tag{4.67}$$

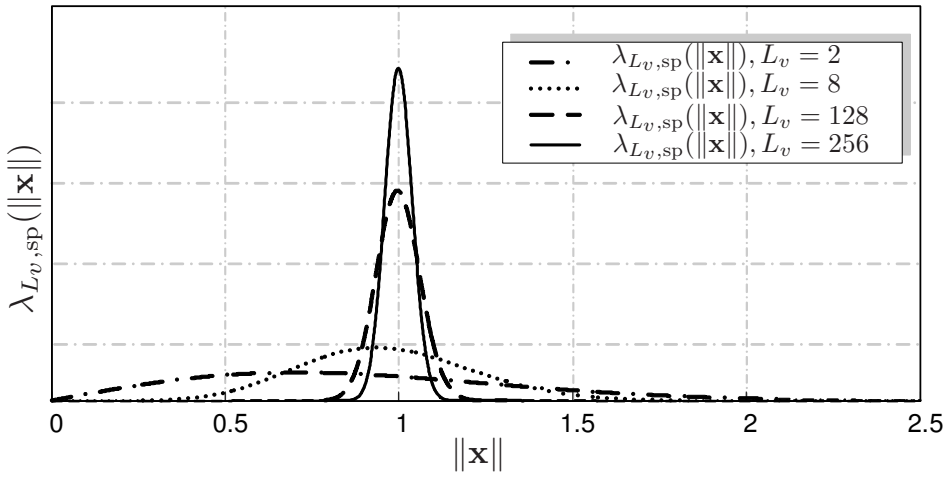


Figure 4.14: (Qualitative) sphere density function computed for the Gaussian PDF for dimensions $L_v = 2$, $L_v = 8$, $L_v = 128$, and $L_v = 256$.

in order to achieve a unit mean absolute value of signal vectors independently from the vector dimension. With increasing dimensions the sphere density function more and more has the characteristic of a peak around $\|x\| = 1.0$ which asymptotically approaches a Dirac. This result is actually well-known as for long analysis segments the short-term power of a stochastic signal tends to be identical to the variance of the signal if stationary. For LSVQ, this effect can be interpreted such that given a stationary Gaussian source and for infinite dimensions, according to the optimal codevector density, all codevectors should be located on the surface of the same unit sphere and hence $N_g = 1$.

4.4 LSVQ Application Examples

In the following, three realizations of LSVQ based on different spherical codes shall be explained and investigated at first. In the second step, measured quantization results will be presented and rated in the context of the theoretical results from Section 4.2.

Numerous contributions have been made on applications of spherical quantization, starting with [Pea79], and [BJ79] where quantization concepts based on a representation in polar coordinates for bivariate circularly symmetric densities were proposed. This concept was generalized to multi-dimensional spherical quantization in polar coordinates in [ST83]. Considering lattice based spherical quantization for speech coding, SVQ realizations were proposed based on the eight-dimensional Gosset Lattice [ALL84], binary and Reed-Muller codes [AL87], the 16-dimensional Barnes-Wall Lattice [LAMM89], and the 24-dimensional Leech Lattice [AB88]. These concepts were part of a hunt for vector quantizers of very high dimension with low bit rates finally resulting in the algebraic codebooks already mentioned at the beginning of this chapter. Outside the speech coding community, applications of lattice based SVQ have also been investigated, e.g., in [SG93] where the Gosset

Lattice is proposed as one application of SVQ for the purpose of image sequence coding. As an additional output of the theoretical work of Hamkins [Ham96], SVQ based on two spherical codes, denoted as Wrapped [HZ97a] and Laminated [HZ97b] Lattices, was proposed.
Besides lattice based approaches contributions have also been made on unstructured spherical codes. For example in [GHSW87], the design of a spherical code based on a *simulated annealing* approach is described for the purpose of channel coding. As an appendix of the referenced paper, the *Apple Peeling* spherical code was introduced. A novel application of the Apple Peeling principle for SVQ for source coding is given in [KV06b] and [KV06a]. A very similar principle is the basis for another approach for SVQ, denoted as Spherical Logarithmic Quantization (SLQ), as proposed in [HM04], [MBH06], and [Mat07].
In the following section, three realizations of SVQ will be presented: The first candidate, SVQ (A), is based on a generalization of the Gosset Lattice. It achieves a very high quantization performance for memoryless sources and dimensions which are multiples of eight. The second candidate, SVQ (B), is a generalization of the algebraic codebooks and the basis for various standardized speech codecs designed for very low bit rates. The last candidate, SVQ(C), is based on the mentioned Apple Peeling code and the most flexible approach. This candidate is also well suited for the quantization of correlated signals with low computational complexity (see Section 6.1).
New computationally efficient nearest neighbor codevector search procedures and index to codevector mapping algorithms will be proposed in the following and in Section D of the *supplement document* [Krü09]. Therefore, all three approaches have a high relevance for applications of source coding in practice. The different candidates for SVQ principally can be combined with logarithmic (A-Law) SQ of the gain factor following the parallel, the sequential, or the joint approach as described in Section C of the *supplement document* [Krü09].
In this chapter only the basic principle of the approaches SVQ (A) and SVQ (B) shall be presented. More details on these approaches and the algorithms for nearest neighbor quantization are provided in Section D of the *supplement document* [Krü09].

4.4.1 SVQ (A): Gosset Low Complexity Vector Quantizer (GLCVQ)

The Gosset Low Complexity Vector Quantizer (GLCVQ) is based on codevectors taken from a generalized version of the eight-dimensional Gosset Lattice E_8 which was at first described in [Gos00]. In contrast to other approaches from the literature based on the Gosset Lattice, the GLCVQ is applicable in various dimensions. Novel nearest neighbor quantization procedures which can be realized with very low complexity and memory consumption even for higher bit rates and dimensions are in detail explained in [KGV08] and Section D.2 of the *supplement document* [Krü09].

The Gosset Lattice is known to have the highest possible density of vectors in eight dimensions (providing a solution for the *Kissing Number Problem* [Bli35]). In order not to be restricted to this dimensionality, the GLCVQ is based on the generalization of the Gosset Lattice for arbitrary dimensions.
The GLCVQ is principally based on a lattice (Lattice VQ, LVQ) but, as a candidate for SVQ, it can be efficiently realized as a permutation code VQ (PCVQ). Both, LVQ and PCVQ, are more in detail explained in Section B of the *supplement document* [Krü09].

4.4.1.1 The GLCVQ Spherical Codebook

The Gosset Lattice is defined as the superposition of the checkerboard lattice D_{L_v} and a shifted version thereof for $L_v = 8$,

$$E_{L_v} := D_{L_v} \cup \Big(D_{L_v} + \mathbf{v} \Big), \mathbf{v} = \begin{bmatrix} \frac{1}{2} & \dots & \frac{1}{2} \end{bmatrix}^T . \tag{4.68}$$

The checkerboard lattice is defined as

$$D_{L_v} := \left\{ \mathbf{x} = \begin{bmatrix} x_0 & \dots & x_{L_v-1} \end{bmatrix}^T \in \mathbb{Z}^{L_v} : (\sum_{i=0}^{L_v-1} x_i) \text{mod } 2 \equiv 0 \right\}. \tag{4.69}$$

Both formulas (4.68) and (4.69) can be easily generalized for arbitrary dimensions L_v. Lattice vectors with a constant distance to the origin define a *shell of a lattice*, $\mathcal{S}$. Different shells of the Gosset Lattice are labeled $\mathcal{S}_{N_{(A)}}^{(E_{L_v})}$ with the code construction parameter $N_{(A)} = 1, 2, \dots \in \mathbb{N}$ as the shell index. For L_v being a multiple of eight, the shell radius is $r_{N_{(A)}}^{(E_{L_v})} = \sqrt{2 \cdot N_{(A)}}$ and the vectors located on each shell are defined as

$$\mathcal{S}_{N_{(A)}}^{(E_{L_v})} := \{ \mathbf{x} \in E_{L_v} : \|\mathbf{x}\| = r_{N_{(A)}}^{(E_{L_v})} \}. \tag{4.70}$$

A *Gosset Lattice spherical codebook*, $\tilde{\mathcal{X}}_{\text{svq},N_{(A)}}^{(E_{L_v})}$, is composed of all vectors related to a specific shell with index $N_{(A)}$, normalized to have unit absolute value,

$$\tilde{\mathcal{X}}_{\text{svq},N_{(A)}}^{(E_{L_v})} := \{ \tilde{\mathbf{c}} = \frac{\mathbf{x}}{\sqrt{2 \cdot N_{(A)}}} \quad \forall \quad \mathbf{x} \in \mathcal{S}_{N_{(A)}}^{(E_{L_v})} \} \tag{4.71}$$

An example for shells of a lattice for the D_2 lattice is shown in Figure 4.15. The number of codevectors located on a shell of the E_{L_v} lattice is

$$N_{\text{svq},N_{(A)}}^{(E_{L_v})} = |\tilde{\mathcal{X}}_{\text{svq},N_{(A)}}^{(E_{L_v})}| \tag{4.72}$$

and is a function of $N_{(A)}$. Examples for the number of codevectors for different values of $N_{(A)}$ for the E_{16} lattice are listed in Table 4.1 together with the effective bit rate

$$R_{\text{eff,svq},N_{(A)}}^{(E_{16})} = \frac{1}{16} \cdot \log_2(N_{\text{svq},N_{(A)}}^{(E_{16})}) \tag{4.73}$$

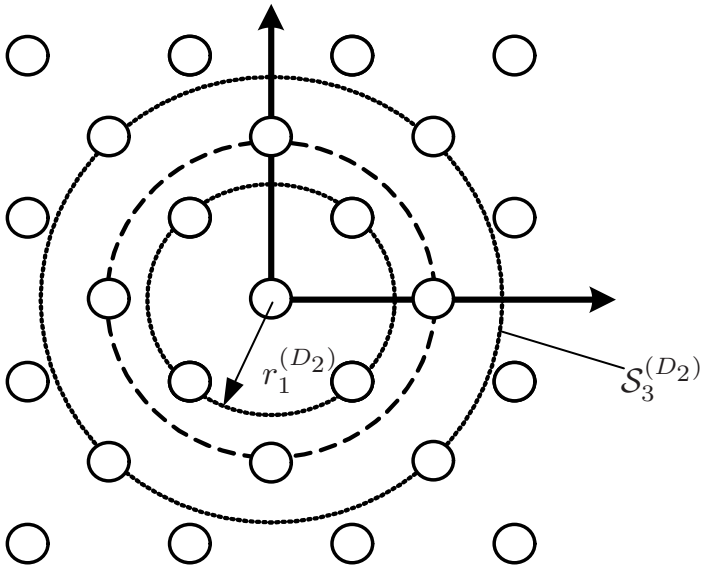


Figure 4.15: Example to illustrate three shells of the D_2 lattice. The radius of the inner most shell is labeled as $r_1^{(D_2)}$, the third shell as $\mathcal{S}_3^{(D_2)}$.

Shell index, $N_{(A)}$	Number of codevectors, $N_{\text{svq},N_{(A)}}^{(E_{16})}$	Effective bit rate per coordinate, $R_{\text{eff, svq},N_{(A)}}^{(E_{16})}$
1	480	0.556
2	61920	0.994
3	1050240	1.25
4	7926240	1.432
5	37500480	1.572
6	135480960	1.688
7	395301120	1.784
8	1014359200	1.86
9	2296875360	1.943
10	4837451920	2.02
15	82051050240	2.26
20	619245426240	2.44

Table 4.1: Codebooks related to different shells of the 16-dimensional (generalized) Gosset Lattice E_{16}.

in bits per vector coordinate to address all spherical codevectors for $L_v = 16$.

Note that if the vector dimension is not multiples of eight, the vectors produced in the context of the checkerboard lattice (first part of (4.68)) and those produced by

the shifted version of the checkerboard lattice (second part of (4.68)) are not located on shells with the same radius (4.70). In this case, a codebook can be determined but the codevectors are not from the Gosset Lattice but from the checkerboard lattice D_{L_v} instead, and the resulting quantization performance is somewhat lower. Also, codevectors are not necessarily located on all shells with the radius according to (4.70) for lower dimensions.
The GLCVQ has recently been proposed as part of the speech and audio codec that has been submitted to ITU-T by Huawei and ETRI as a candidate for the upcoming super-wideband and stereo extensions of Rec. G.729.1 and G.718 [GKL+09].

4.4.2 SVQ (B): Algebraic Low Bit Rate Vector Quantizer (ALBVQ)

The algebraic low bit rate vector quantizer (ALBVQ) is a modification of the vector codebooks employed in speech coding known as the *algebraic codebooks*. Those were designed principally for very low bit rates to quantize the linear prediction residual signal in linear predictive coding [LASM90].
In order to be efficiently applicable for higher bit rates and for quantization in general, new nearest neighbor quantization procedures and codeword-to-codevector mapping algorithms are proposed and in detail explained in Section D.3 of the *supplement document* [Krü09]). These new techniques deviate from those known from the literature. In this section only the basic principle of the ALBVQ vector quantizer will be briefly summarized. Very similar to the GLCVQ, the high coding efficiency of this approach is based on the fact that the ALBVQ can be realized as a permutation code VQ (refer to Section B.2 of the *supplement document* [Krü09]).

4.4.2.1 The ALBVQ Spherical Codebook

The algebraic codebook is defined as the amount of codevectors that can be constructed by setting a specific number of ternary pulses at arbitrary positions. In practice, the number of ternary pulses which have non-zero amplitudes is often very low, therefore these codebooks are often referred to as *sparse codebooks*. An example codevector of length $L_v = 16$ is shown in Figure 4.16. Among all available positions, at positions $\nu_{\text{pos},0} = 4$ and $\nu_{\text{pos},1} = 7$ ternary pulses are set to the amplitude of $x_{\nu_{\text{pos},0}} = +1$ and $x_{\nu_{\text{pos},1}} = -1$. The number of non-zero constant amplitude pulses is the code construction parameter $N_{(B)}$ with $N_{(B)} = 2$ in the example in the figure. Among the L_v positions in the pulse train, not all are necessarily allowed. Therefore $\mathcal{I}_{\text{pos}}$ is defined as the amount of valid positions with

$$\nu_{\text{pos},\kappa} \in \mathcal{I}_{\text{pos}} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\} \tag{4.74}$$

in the example of the figure to allow all positions. If all positions are allowed with a maximum of one pulse to be set at each position[8], valid codevectors are defined

[8] This restriction is not necessarily given in standardized speech codecs which are based on algebraic codebooks.

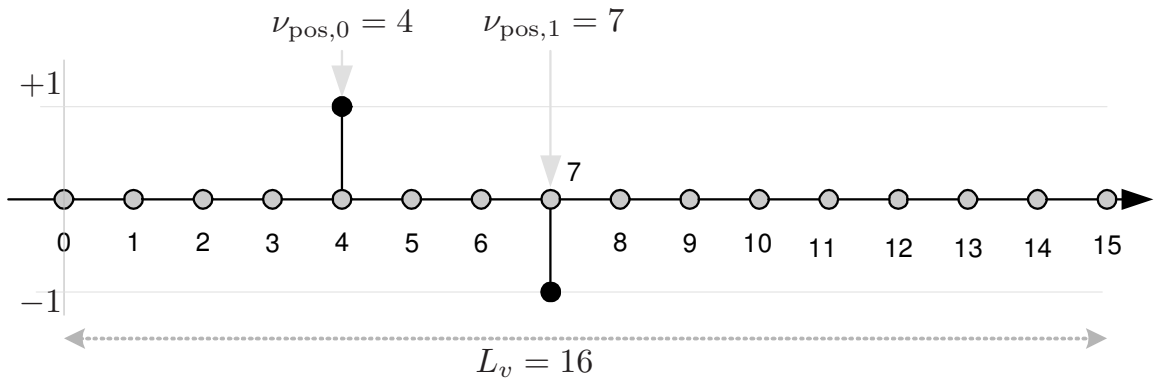


Figure 4.16: Algebraic codevector: Example with ternary pulses at two positions.

as

$$\mathcal{S}_{N_{(B)}}^{(\mathcal{A}_{L_v})} := \left\{ \mathbf{x} = \begin{bmatrix} x_0 & \dots & x_{L_v-1} \end{bmatrix}^T \in \mathbb{Z}^{L_v} : x_i \in \{-1, 0, 1\} : \sum_{i=0}^{L_v-1} |x_i| = N_{(B)} \right\}. \tag{4.75}$$

In this case all valid codevectors $\mathbf{x}$ are located on the surface of a sphere with radius $r_{N_{(B)}}^{(\mathcal{A}_{L_v})} = \sqrt{N_{(B)}}$. The *spherical codebook* $\tilde{\mathcal{X}}_{\text{svq},N_{(B)}}^{(\mathcal{A}_{L_v})}$ is composed of all vectors in $\mathcal{S}_{N_{(B)}}^{(\mathcal{A})}$, normalized to have unit absolute value,

$$\tilde{\mathcal{X}}_{\text{svq},N_{(B)}}^{(\mathcal{A}_{L_v})} := \{ \tilde{\mathbf{c}} = \frac{\mathbf{x}}{\sqrt{N_{(B)}}} \quad \forall \quad \mathbf{x} \in \mathcal{S}_{N_{(B)}}^{(\mathcal{A}_{L_v})} \} \tag{4.76}$$

The number of codevectors depends on the vector dimension L_v and the number $N_{(B)}$ of ternary pulses to be set. In Table 4.2 this number is listed together with the corresponding effective bit rate

$$R_{\text{eff,svq},N_{(B)}}^{(\mathcal{A}_{16})} = \frac{1}{16} \cdot \log_2(N_{\text{svq},N_{(B)}}^{(\mathcal{A}_{16})}) \tag{4.77}$$

in bits per vector coordinate for $L_v = 16$ to address all codevectors and different values of $N_{(B)}$. In contrast to the GLCVQ where the number of codevectors more or less monotonically increases for increasing values of the code construction parameter $N_{(A)}$, in this case, the number of codevectors increases with increasing values of $N_{(B)}$, then reaches a maximum number and finally decreases to reach an effective bit rate of $R_{\text{eff,svq},N_{(B)}}^{(\mathcal{A}_{L_v})} = 1$ if $N_{(B)} = L_v$. In this case, only the signs of the pulses at all position are coded.

4.4.3 SVQ (C): Apple Peeling Vector Quantizer (APVQ)

The Apple Peeling spherical code was introduced originally as a channel code in [GHSW87]. In order to employ the same spherical code for VQ, efficient quantization algorithms (also in combination with linear prediction), codebook representations, and index-to-codeword mapping algorithms have been proposed in [KV06b]

Number pulses, $N_{(B)}$	Number of codevectors, $N_{\text{svq},N_{(B)}}^{(\mathcal{A}_{16})}$	Effective bit rate per coordinate, $R_{\text{eff},\text{svq},N_{(B)}}^{(\mathcal{A}_{16})}$
1	32	0.3125
2	480	0.5566
3	4480	0.758
4	29120	0.9268
6	512512	1.18
8	3294720	1.353
10	8200192	1.435
12	7454720	1.426
14	1966080	1.3
15	524288	1.187
16	65536	1

Table 4.2: Number of codevectors for the 16-dimensional ALBVQ as a function of the number of ternary pulses.

and [KV06a]. The resulting LSVQ is denoted as the *Apple Peeling Vector Quantizer* (APVQ). Due to its high coding efficiency, the APVQ is also the basis for the SCELP and the W-SCELP low delay audio codecs which will be described in Chapter 6. The principle of the APVQ shall at first be explained based on the example for $L_v = 3$ and will then be generalized for arbitrary dimensions afterwards.
A specific nearest neighbor quantization procedure as for SVQ(A) and (B) does not exist. Instead, a low complexity quantization procedure for memoryless sources can be derived from the weighted vector search explained in the context of CELP coding in Section E.1 of the *supplement document* [Krü09].
The APVQ can be realized by means of *history specific* uniform quantization in polar coordinates. The term *history specific* indicates that different uniform scalar quantizers are employed for each polar coordinate depending on the *history* of previously quantized polar coordinate(s). The impact of the uniform scalar quantizers for all polar coordinates will be analyzed in the context of the resulting overall spherical vector codebook $\tilde{\mathcal{X}}_{\text{svq},N_{(C)}}^{(\text{AP}_{L_v})}$ in the following.

4.4.3.1 Cartesian to Polar Coordinate Transform

The Apple Peeling spherical code is related to a representation of vectors in polar coordinates. Therefore, the cartesian to polar coordinate transform plays an important role for the construction of the APVQ codevectors and is shortly reviewed. A vector given in *cartesian coordinates*,

$$\mathbf{x}_{\text{cartesian}} = \begin{bmatrix} x_0 & x_1 & \dots & x_{L_v-1} \end{bmatrix}^T, \tag{4.78}$$

is written in *polar coordinates* as

$$\mathbf{x}_{\text{polar}} = \begin{bmatrix} r_{\mathbf{x}} & \vartheta_{\mathbf{x},0} & \dots & \vartheta_{\mathbf{x},(L_v-2)} \end{bmatrix}^T . \tag{4.79}$$

The polar coordinates based representation is composed of a radius $r_{\mathbf{x}}$ and $(L_v - 1)$ angles $\vartheta_{\mathbf{x},\nu}$. Given the normalized vector $\mathbf{c}$ from (4.6), the radius is $r_{\mathbf{c}} = 1$. Since the spherical codevectors to be constructed in the following are all located on the surface of a unit sphere, we will mainly focus on the angles $\vartheta_{\mathbf{c},\nu}$ for normalized vectors $\mathbf{c}$ in the codevector definition.
In order to cover the complete surface of a unit sphere in L_v dimensions, the angles are in the range of

$$\vartheta_{\mathbf{c},\nu} \in \begin{cases} [0,\pi] & \text{for } \nu \neq 0 \\ [0,2\pi) & \text{for } \nu = 0 \end{cases} \tag{4.80}$$

Given a polar representation of a vector $\mathbf{c}$ and thus angles $\vartheta_{\mathbf{c},\nu}$, in order to reconstruct all coordinates of vector $\mathbf{c}$ in cartesian coordinate representation, a recursive algorithm to calculate all cartesian coordinates for decreasing indices $\nu = (L_V - 2), \dots, 0$ can be written as (e.g. [Coo52]):

$$\begin{aligned} c_{L_v-2-\nu} &= \cos(\vartheta_{\mathbf{c},\nu}) \cdot r_{\mathbf{c},\nu+2} \\ r_{\mathbf{c},\nu+1} &= \sin(\vartheta_{\mathbf{c},\nu}) \cdot r_{\mathbf{c},\nu+2}. \end{aligned} \tag{4.81}$$

Finally, the coordinate with index $(L_v - 1)$ is reconstructed as

$$c_{L_v-1} = r_{\mathbf{c},1}. \tag{4.82}$$

Since vector $\mathbf{c}$ has unit length, the start value for the radius is

$$r_{\mathbf{c},L_v} = r_{\mathbf{c}} = 1.0. \tag{4.83}$$

This algorithm can be interpreted as follows: According to (4.81), for each ν, a vector of length $r_{\mathbf{c},\nu+2}$ and dimension $\nu + 2$ is decomposed into a height $c_{L_v-2-\nu}$ and a projection of the vector into a plane of dimension $\nu + 2 - 1$ spanned by all remaining coordinates $c_{L_v-2-\nu+1}$, $c_{L_v-2-\nu+2}$, $\dots$, c_{L_v-1}. The height $c_{L_v-2-\nu}$ is calculated in the first, and the length of the projected vector in the second part of (4.81).
In Figure 4.17, an example for the polar to cartesian coordinate transform for a normalized vector $\mathbf{c}$ is shown for $L_v = 3$ with $r_{\mathbf{c},3} = 1.0$ and

$$\begin{aligned} c_0 &= \cos(\vartheta_{\mathbf{c},1}) \cdot r_{\mathbf{c},3} & r_{\mathbf{c},2} &= \sin(\vartheta_{\mathbf{c},1}) \cdot r_{\mathbf{c},3} \\ c_1 &= \cos(\vartheta_{\mathbf{c},0}) \cdot r_{\mathbf{c},2} & r_{\mathbf{c},1} &= \sin(\vartheta_{\mathbf{c},0}) \cdot r_{\mathbf{c},2} \\ c_2 &= r_{\mathbf{c},1} \end{aligned} \tag{4.84}$$

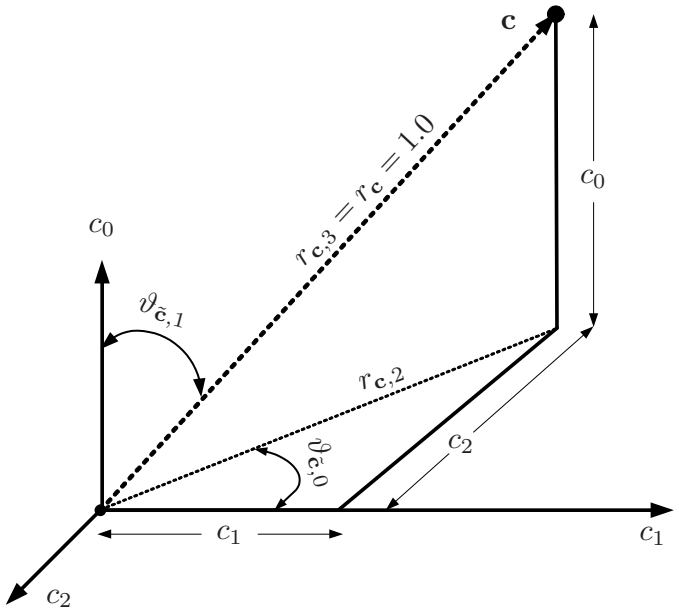


Figure 4.17: Polar to cartesian coordinate transform, example for $L_v = 3$.

Given a normalized vector $\mathbf{c}$ in cartesian coordinates, the recursive algorithm (4.81) can also be used the other way around to calculate all angles in the polar representation for $\nu = (L_V - 2), \ldots, 0$:

$$\begin{aligned} \vartheta_{\mathbf{c},\nu} &= \arccos\left(\frac{c_{L_v-2-\nu}}{r_{\mathbf{c},\nu+2}}\right) \\ r_{\mathbf{c},\nu+1} &= \sin(\vartheta_{\mathbf{c},\nu}) \cdot r_{\mathbf{c},\nu+2} \end{aligned} \tag{4.85}$$

with the start value given in (4.83).

4.4.3.2 The APVQ Spherical Codebook for $L_v = 3$

The construction of the codevectors $\tilde{\mathbf{c}}$ of the APVQ will be shown for the example of $L_v = 3$. The design target of the codevector construction is to place all spherical codevectors uniformly on the surface of the unit sphere. The overall number of spherical codevectors depends on the choice of code construction parameter $N_{(C)}$ to be defined in the following.
For the explanation of the code construction, the upper half of an exemplary unit sphere is shown in Figure 4.18 for $L_v = 3$. It is sufficient to show the upper half of the sphere only because the codevectors are generated symmetrically on the upper and the lower half. In the figure, the angles $\vartheta_{\tilde{\mathbf{c}},0}$ and $\vartheta_{\tilde{\mathbf{c}},1}$ are shown in analogy to Figure 4.17. The big black dots labeled as $\tilde{\mathbf{c}}_a$-$\tilde{\mathbf{c}}_e$ are example spherical codevectors. For the construction of codevectors in three dimensions, at first, the angle $\vartheta_{\tilde{\mathbf{c}},1} \in [0, \pi]$ is uniformly quantized. In this context, the code construction parameter $N_{(C)}$ is defined as the number of reconstruction levels at the positions

$$\tilde{\vartheta}_{\tilde{\mathbf{c}},1}^{(m_{\vartheta_1})} = (m_{\vartheta_1} + \frac{1}{2}) \cdot \frac{\pi}{N_{(C)}} \text{ with } 0 \leq m_{\vartheta_1} < N_{(C)} \tag{4.86}$$

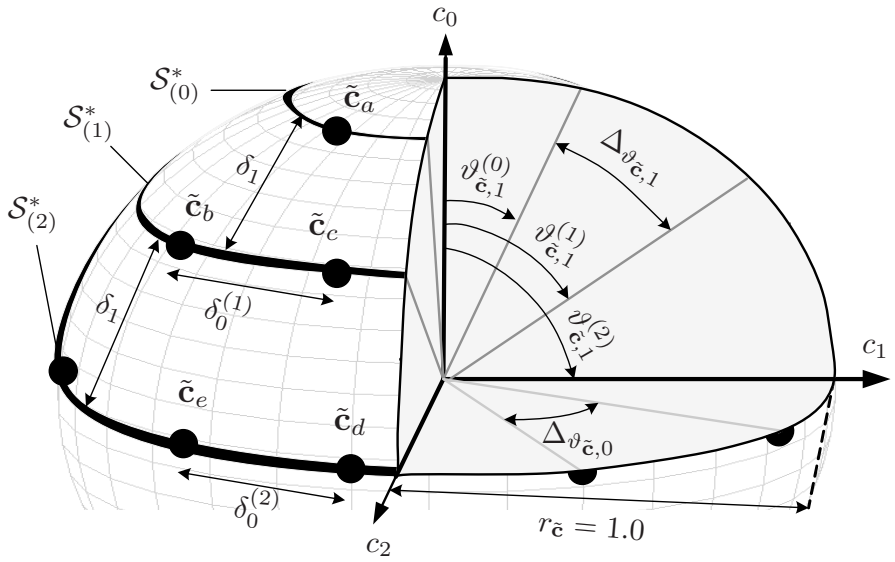


Figure 4.18: Construction of Apple Peeling spherical codevectors, example for $L_v = 3$ and $N_{(C)} = 5$.

yielding constant angle quantization intervals of width

$$\Delta_{\vartheta_{\tilde{\mathbf{c}},1}} = \frac{\pi}{N_{(C)}} \tag{4.87}$$

Considering the transform from polar to cartesian coordinates in (4.81), based on the quantizer reconstruction levels for the first angle also the first cartesian coordinate of the codevectors can be computed as

$$\tilde{c}_0^{(m_{\vartheta_1})} = \cos(\tilde{\vartheta}^{(m_{\vartheta_1})}_{\tilde{\mathbf{c}},1}) \tag{4.88}$$

By quantizing their first cartesian coordinate, all codevectors are restricted to be located on one among $N_{(C)}$ planes which are parallel to the c_1-c_2-plane at heights of $c_0 = \tilde{c}_0^{(m_{\vartheta_1})}$ for $0 \leq m_{\vartheta_1} < N_{(C)}$. Combining these planes with the sphere surface, circles $\mathcal{S}^*_{(m_{\vartheta_1})}$ are the result which are depicted in the figure for $m_{\vartheta_1} = 0, 1, 2$. The closest distance from one circle to its neighbor is δ_1 as shown in the figure,

$$\delta_1 = 2 \cdot \sin(\Delta_{\vartheta_{\tilde{\mathbf{c}},1}}/2) \cdot r_{\tilde{\mathbf{c}}} = 2 \cdot \sin(\Delta_{\vartheta_{\tilde{\mathbf{c}},1}}/2) \approx \Delta_{\vartheta_{\tilde{\mathbf{c}},1}} = \frac{\pi}{N_{(C)}} \tag{4.89}$$

with $r_{\tilde{\mathbf{c}}} = 1$ because all codevectors have unit absolute value. The approximation of the distance by the circular arc is made because $\Delta_{\vartheta_{\tilde{\mathbf{c}},1}}$ is assumed to be very small. The radius $r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})}$ of each circle $\mathcal{S}^*_{(m_{\vartheta_1})}$ is the projection of the overall radius $r_{\tilde{\mathbf{c}}}$ according to the second part of (4.81) and depends on the quantization index m_{ϑ_1}

$$r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})} = \sin(\tilde{\vartheta}^{(m_{\vartheta_1})}_{\tilde{\mathbf{c}},1}) \cdot r_{\tilde{\mathbf{c}}} = \sin(\tilde{\vartheta}^{(m_{\vartheta_1})}_{\tilde{\mathbf{c}},1}) \tag{4.90}$$

In the next step, the angle $\vartheta_{\tilde{\mathbf{c}},0}$ is quantized to $\tilde{\vartheta}_{\tilde{\mathbf{c}},0}$. Again, a uniform quantization is employed here with $N^{(m_{\vartheta_1})}$ as the (yet unknown) number of quantization reconstruction levels which may be different for each of the circles $\mathcal{S}^*_{(m_{\vartheta_1})}$ with index m_{ϑ_1}. With respect to the overall range of angle $\vartheta_{\tilde{\mathbf{c}},0} \in [0, 2\pi)$, the quantization reconstruction levels are

$$\tilde{\vartheta}_{\tilde{\mathbf{c}},0}^{(m_{\vartheta_1},m_{\vartheta_0})} = (m_{\vartheta_0} + \frac{1}{2}) \cdot \frac{2 \cdot \pi}{N^{(m_{\vartheta_1})}}. \tag{4.91}$$

with $0 \leq m_{\vartheta_0} < N^{(m_{\vartheta_1})}$. The corresponding quantization interval width is

$$\Delta_{\vartheta_{\tilde{\mathbf{c}},0}}^{(m_{\vartheta_1})} = \frac{2 \cdot \pi}{N^{(m_{\vartheta_1})}}. \tag{4.92}$$

Given the quantization index m_{ϑ_0}, all remaining coordinates of the spherical codevectors can be calculated in analogy to (4.81) as

$$\begin{aligned} \tilde{c}_1^{(m_{\vartheta_1},m_{\vartheta_0})} &= \cos(\tilde{\vartheta}_{\tilde{\mathbf{c}},0}^{(m_{\vartheta_1},m_{\vartheta_0})}) \cdot r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})} \\ \tilde{c}_2^{(m_{\vartheta_1},m_{\vartheta_0})} &= \sin(\tilde{\vartheta}_{\tilde{\mathbf{c}},0}^{(m_{\vartheta_1},m_{\vartheta_0})}) \cdot r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})}. \end{aligned} \tag{4.93}$$

Each spherical codevector is identified uniquely by a tuple of indices $\begin{bmatrix} m_{\vartheta_1} & m_{\vartheta_0} \end{bmatrix}$. Note that the variable $N^{(m_{\vartheta_1})}$ is still unknown. In order to compute it, the quantization interval $\Delta_{\vartheta_{\tilde{\mathbf{c}},0}}^{(m_{\vartheta_1})}$ is transformed into a distance on the sphere surface between adjacent codevectors on each circle, $\delta_0^{(m_{\vartheta_1})}$. In analogy to (4.89), this distance can be approximated by the circular arc

$$\delta_0^{(m_{\vartheta_1})} \approx \Delta_{\vartheta_{\tilde{\mathbf{c}},0}}^{(m_{\vartheta_1})} \cdot r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})} = \frac{2 \cdot \pi}{N^{(m_{\vartheta_1})}} \cdot r_{\tilde{\mathbf{c}},2}^{(m_{\vartheta_1})} \tag{4.94}$$

Recalling the global target to distribute codevectors uniformly over the surface of the sphere, the *Apple Peeling constraint* specifies that the distance between circles (4.89) shall be identical to the distance between the codevectors on each circle (4.94),

$$\delta_0^{(m_{\vartheta_1})} \stackrel{!}{=} \delta_1 \approx \frac{\pi}{N_{(C)}}, \tag{4.95}$$

exemplified in Figure 4.18. Since $N^{(m_{\vartheta_1})}$ must be an integer, it is computed as

$$N^{(m_{\vartheta_1})} = \lfloor 2 \cdot N_{(C)} \cdot \sin(\tilde{\vartheta}_{\tilde{\mathbf{c}},1}^{(m_{\vartheta_1})}) \rfloor. \tag{4.96}$$

and is a function of the *quantization history* which is the quantized angle $\tilde{\vartheta}_{\tilde{\mathbf{c}},1}^{(m_{\vartheta_1})}$ here.

4.4.3.3 Generalization of the APVQ for Arbitrary Dimensions

In order to generalize the concept described for $L_v = 3$ to arbitrary dimensions, the aggregation of indices for previously quantized angles is written as one vector $\boldsymbol{\Xi}_\nu$ (the *quantization history*) for the purpose of simplification of the notation,

$$\boldsymbol{\Xi}_\nu = \begin{bmatrix} m_{\vartheta_{L_v-2}} & m_{\vartheta_{L_v-3}} & \dots & m_{\vartheta_\nu} \end{bmatrix}^T. \tag{4.97}$$

In addition, a partial codevector is defined as

$$\tilde{\mathbf{c}}^{(\boldsymbol{\Xi}_\nu)} = \begin{bmatrix} \tilde{c}_0^{(\boldsymbol{\Xi}_\nu)} & \tilde{c}_1^{(\boldsymbol{\Xi}_\nu)} & \dots \tilde{c}_{(L_v-1-\nu)}^{(\boldsymbol{\Xi}_\nu)} \end{bmatrix}^T. \tag{4.98}$$

For the construction of the APVQ codevectors, a recursive algorithm is defined in Figure 4.19 with the start values

$$\begin{aligned} \nu &= (L_v - 2) \\ \boldsymbol{\Xi}_{\nu+1} &= \boldsymbol{\Xi}_{(L_v-1)} = [\,] \\ r_{\tilde{\mathbf{c}},\nu+2}^{(\boldsymbol{\Xi}_{\nu+1})} &= r_{\tilde{\mathbf{c}},L_v}^{(\boldsymbol{\Xi}_{(L_v-1)})} = r_{\tilde{\mathbf{c}}} = 1.0 \\ \tilde{\mathbf{c}}^{(\boldsymbol{\Xi}_{\nu+1})} &= \tilde{\mathbf{c}}^{(\boldsymbol{\Xi}_{(L_v-1)})} = [\,]^T \\ N^{(\boldsymbol{\Xi}_{\nu+1})} &= N^{(\boldsymbol{\Xi}_{(L_v-1)})} = N_{(C)}. \end{aligned} \tag{4.99}$$

The described procedure can also be interpreted in the context of a *code construction tree* as shown in Figure 4.20 for the example of $L_v = 3$. Each node in the tree (except for the *root*) corresponds to one step of the recursive algorithm. From each node, each branch represents one index value m_{ϑ_ν}. The copy and update functionality is shown for the history $\boldsymbol{\Xi}_\nu$ and the partial codevector $\mathbf{c}^{(\boldsymbol{\Xi}_\nu)}$ for two example nodes. Each of the leafs of the codevector tree corresponds to one spherical codevector. This tree representation of the codevector construction will also be reviewed for the development of the efficient quantization procedure described in Section E.1 of the *supplement document* [Krü09]. Note that for an efficient employment of the APVQ for quantization, a compact version of the overall vector codebook must be stored based on a technique published in [KV06a]. The proposed technique enables an exemplary reduction of the required read-only-memory (ROM) by a factor of approximately 1390 compared to conventional lookup tables.

Given a value $N_{(C)}$, the number of spherical codevectors $N_{\mathrm{svq},N_{(C)}}^{(\mathrm{AP}_{L_v})}$ cannot be calculated analytically. In Table 4.3, for the example of $L_v = 16$ and different values of $N_{(C)}$, the number of spherical codevectors is presented together with the effective bit rate

$$R_{\mathrm{eff,svq},N_{(C)}}^{(\mathrm{AP}_{16})} = \frac{1}{16} \cdot \log_2(N_{\mathrm{svq},N_{(C)}}^{(\mathrm{AP}_{16})}) \tag{4.114}$$

in bits per vector coordinate.

Recursive Algorithm to Generate Apple Peeling Codevectors

1. Produce the indices for the quantization of the angle $\vartheta_{\tilde{\mathbf{c}},\nu}$,

$$m_{\vartheta_\nu} = 0 \ldots (N^{(\Xi_{\nu+1})} - 1). \tag{4.107}$$

2. For each index m_{ϑ_ν},

- create an updated copy of the quantization history $\Xi_{\nu+1}$,

$$\Xi_\nu = \begin{bmatrix} \Xi_{\nu+1} & m_{\vartheta_\nu} \end{bmatrix}, \tag{4.108}$$

- produce the quantized angles (uniform SQ) in analogy to (4.86)

$$\tilde{\vartheta}_{\tilde{\mathbf{c}},\nu}^{(\Xi_\nu)} = (m_{\vartheta_\nu} + \frac{1}{2}) \cdot \frac{1}{N^{(\Xi_{\nu+1})}} \cdot \begin{cases} 2 \cdot \pi & \text{if } \nu = 0 \\ \pi & \text{if } \nu \neq 0 \end{cases}, \tag{4.109}$$

- calculate the next cartesian coordinate and create an updated copy of the partial codevector $\mathbf{c}^{\Xi_\nu+1}$

$$\begin{aligned} \tilde{c}_{L_v-2-\nu}^{(\Xi_\nu)} &= \cos(\tilde{\vartheta}_{\tilde{\mathbf{c}},\nu}^{(\Xi_\nu)}) \cdot r_{\mathbf{c},\nu+2}^{(\Xi_{\nu+1})} \\ \tilde{\mathbf{c}}^{(\Xi_\nu)} &= \begin{bmatrix} \tilde{\mathbf{c}}^{(\Xi_{\nu+1})} & \tilde{c}_{L_v-2-\nu}^{(\Xi_\nu)} \end{bmatrix}^T, \end{aligned} \tag{4.110}$$

- update the radius for each projection onto a sub sphere in analogy to (4.90),

$$r_{\tilde{\mathbf{c}},\nu+1}^{(\Xi_\nu)} = \sin(\tilde{\vartheta}_{\tilde{\mathbf{c}},\nu}^{(\Xi_\nu)}) \cdot r_{\tilde{\mathbf{c}},\nu+2}^{(\Xi_{\nu+1})}, \tag{4.111}$$

- update the number of quantization reconstruction levels for the next angle in analogy to (4.96)($N^{(\Xi_0)}$ is not required),

$$N^{(\Xi_\nu)} = \begin{cases} \lfloor 2 \cdot N^{(\Xi_{\nu+1})} \cdot \sin(\tilde{\vartheta}_{\tilde{\mathbf{c}},\nu}^{(\Xi_\nu)}) \rfloor & \text{if } \nu = 1 \\ \lfloor N^{(\Xi_{\nu+1})} \cdot \sin(\tilde{\vartheta}_{\tilde{\mathbf{c}},\nu}^{(\Xi_\nu)}) \rfloor & \text{if } \nu > 1 \end{cases}, \tag{4.112}$$

- **if** $\nu > 0$: for each copy of the history and the partial codevector, restart the algorithm in 1. with $\nu = \nu - 1$ (reduce the dimension by one),
  if $\nu = 0$: in analogy to (4.84), compute the last coordinate to finalize one additional codevector $\tilde{\mathbf{c}}$ in the codebook,

$$\tilde{\mathbf{c}} = \begin{bmatrix} \tilde{\mathbf{c}}^{(\Xi_0)} & r_{\tilde{\mathbf{c}},1}^{(\Xi_0)} \end{bmatrix}^T. \tag{4.113}$$

Figure 4.19: Update step for the recursive algorithm to generate APVQ codevectors for arbitrary dimensions.

4.4.4 Measured Results

Figures to demonstrate the properties and achievable quantization performance of the three example LSVQ realizations shall be shown in the following. All presented SNR plots are based on simulations in which (quasi) stationary memoryless i.i.d. Gaussian signals were quantized, and the output SNR values were measured for different effective bit rates.

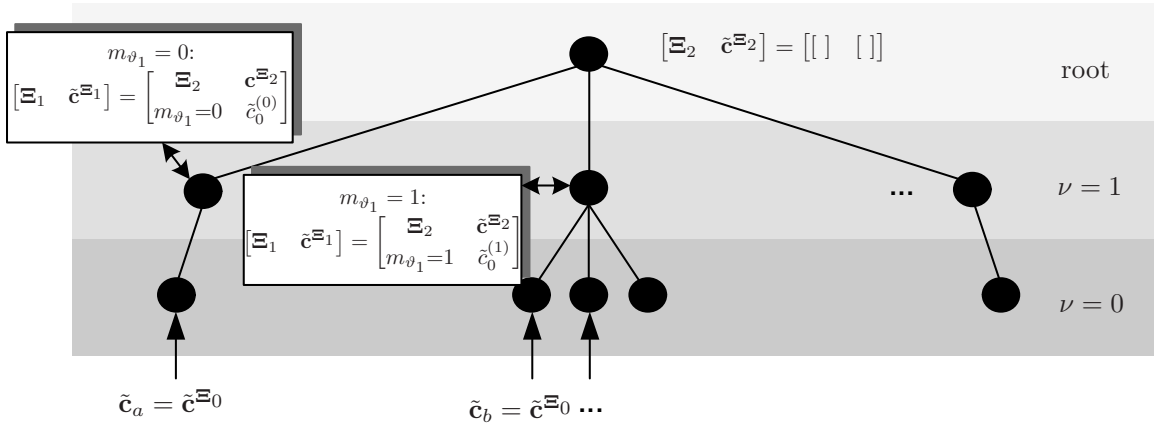


Figure 4.20: Apple Peeling codevector construction tree.

Start parameter, $N_{(C)}$	Number of codevectors, $N_{\text{svq},N_{(C)}}^{(\text{AP}_{16})}$	Effective bit rate per coordinate, $R_{\text{eff,svq},N_{(C)}}^{(\text{AP}_{16})}$
1	2	0.0625
2	4	0.125
3	90	0.40
4	176	0.46
6	2502	0.7
8	1142908	1.25
10	42707208	1.58
11	231267586	1.73
13	3074430078	1.96

Table 4.3: Number of APVQ codevectors for $L_v = 16$ and different values of start parameter $N_{(C)}$.

4.4.4.1 Achievable Bit Rates

In the three approaches, the number of codevectors and hence the effective bit rate $R_{\text{eff,svq}}$ related to Q_{svq} depends on the code construction parameters $N_{(A)}$, $N_{(B)}$, or $N_{(C)}$, respectively. The achievable effective bit rate of each proposed approach controls the quantization quality and hence its applicability for different application types. In Figure 4.21 the achievable effective bit rates are shown for the example of $L_v = 16$ over the code construction parameter $N_{(A)}$, $N_{(B)}$, or $N_{(C)}$.
The GLCVQ can be operated only at bit rates higher than $R_{\text{eff,svq}} > 0.5$ bits per vector coordinate. With increasing values of $N_{(A)}$, also the bit rate increases. The ALBVQ can be operated only at low bit rates. The maximum bit rate is achieved for $N_{(B)} = 10$ for $L_v = 16$. The APVQ is capable to generate codebooks for low and for high bit rates and is also the most flexible approach.

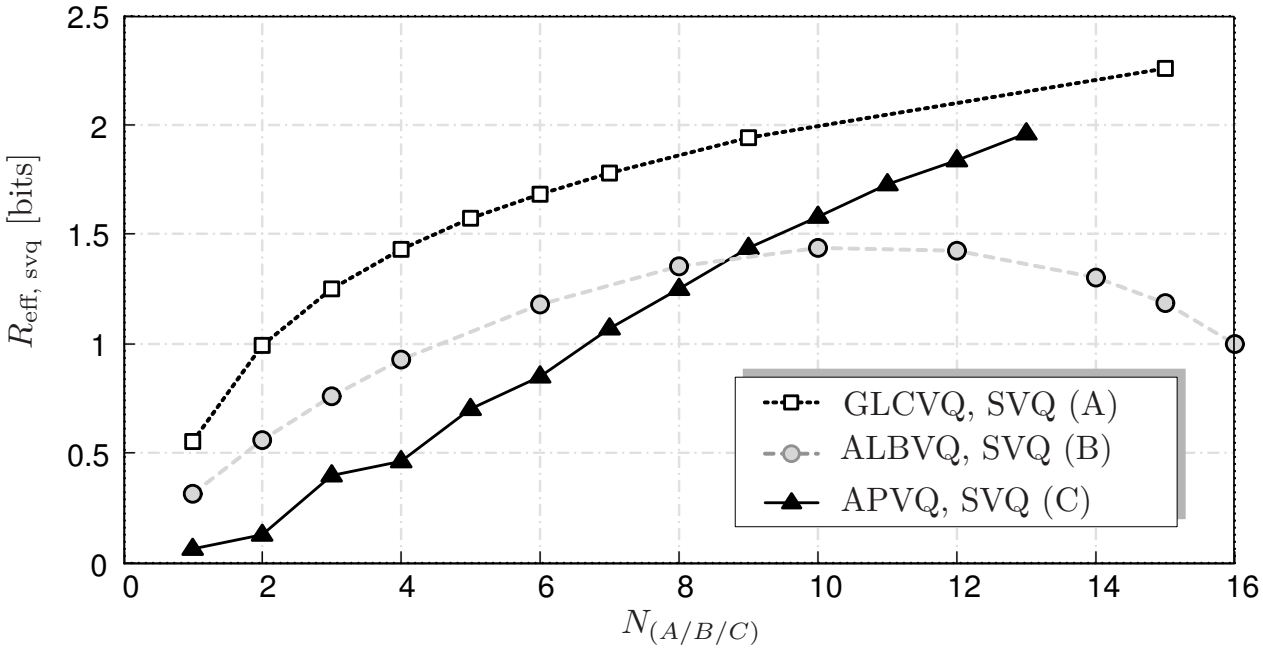


Figure 4.21: Effective bit rate $R_{\text{eff,svq}}$ over code construction parameter $N_{(A/B/C)}$ for $L_v = 16$.

4.4.4.2 SNR Plots related to SVQ

At first, only the SVQ part of the three proposed approaches for LSVQ is assessed. Therefore, the SNR solely related to the quantization of normalized vectors $\mathbf{c}$ (see Figure 4.3) is measured in simulations with different effective bit rates $R_{\text{eff, svq}}$. The results are shown in Figure 4.22 for $L_v = 16$. Note that the SNR was not evaluated for all values of the start parameters $N_{(A)}$, $N_{(B)}$, and $N_{(C)}$. In addition to the measured curves, the theoretical SNR according to (4.33), $\text{SNR}_{\text{svq}}^{(II)}$, is shown as a reference. The performance related to the GLCVQ is closest to the theoretical bound. The ALBVQ also has a high performance for low values of $N_{(B)}$. A very interesting fact is that in case of the lowest bit rate the codevectors related to the GLCVQ and the ALBVQ are identical. For higher values of $N_{(B)}$, however, the SNR for the ALBVQ displaces from the theoretical bound. The maximum bit rate and SNR is finally reached for $N_{(B)} = 8$. The APVQ achieves slightly lower SNR values than the GLCVQ. Very similar results are observed for $L_v = 8$.

4.4.4.3 SNR Plots related to LSVQ

SNR values for LSVQ based on the three candidate SVQ approaches are shown in Figure 4.23 for $L_v = 16$. For each of the candidates, an optimal bit allocation was computed on the basis of (4.53), the A-Law constant was set to $A = 5000$ in analogy to Section 4.3. In addition to the measured values, the theoretical value from (4.43) is depicted as $\text{SNR}_{\text{lsvq}}^{(II)}|_{\text{dB}}$. Also, the asymptotic performance for $L_v \to \infty$, and $\text{SNR}_{Q,\text{nu,A}}|_{\text{dB}}$ as LSVQ for $L_v = 1$ are shown as reference curves.

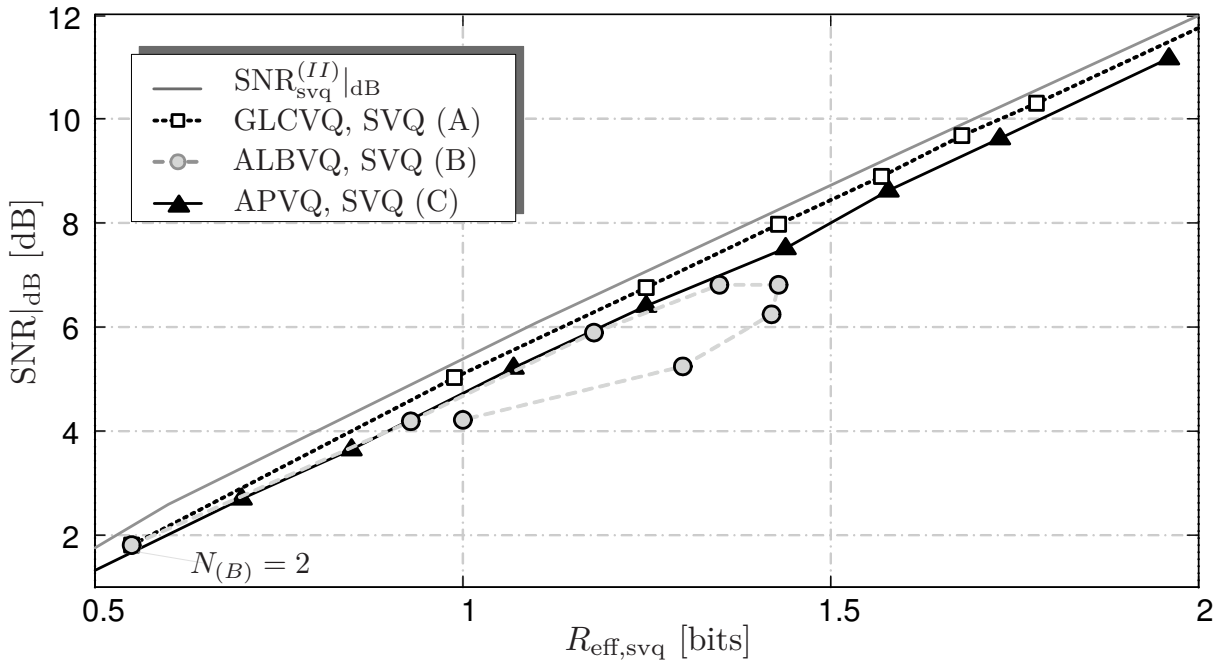


Figure 4.22: SNR for SVQ of normalized vectors $\mathbf{c}$ for $L_v = 16$ and different effective bit rates for the GLCVQ, the ALBVQ, and the APVQ. The curve for the ALBVQ is depicted for $N_{(B)} = 2, 4, 6, 8, 10, 12, 14, 15$ (starting for $N_{(B)} = 2$ for the left most point).

4.4.4.4 GLCVQ versus Leech Lattice SVQ

The *Leech Lattice* [Lee64], [Lee67] is known to have the highest codevector density in 24 dimensions. An approach for SVQ based on the Leech Lattice was proposed in [AB88]. The disadvantage of the described approach is that the nearest neighbor quantization procedures proposed in the Literature are too complex for practical applications. Naturally, even though it can be adapted for 24 dimensions, the GLCVQ approach can not reach the same performance. Nevertheless, SNR values were measured for the GLCVQ for $L_v = 24$ to see how well the GLCVQ performs. The result is shown in Figure 4.24. The reference curve for the Leech Lattice SVQ is based on the results given in [AB88] for a memoryless Gaussian source. The comparison measured for a Gaussian source is somewhat unfair since the Leech Lattice SVQ is combined with a source optimized gain SQ to adapt to the special case of a Gaussian source whereas logarithmic gain SQ is used in the context of the GLCVQ to be applicable for sources with unknown PDFs. In both cases instead of the parallel combination of the quantizers for the shape and the gain components, a sequential approach (refer to Section C of the *supplement document* [Krü09]) is followed. In addition to the measured SNR curves, the theoretical LSVQ bound $\text{SNR}_{\text{lsvq}}^{(II)}|_{\text{dB}}$ for $L_v = 24$ is shown in the figure.
The curves measured for the Leech Lattice SVQ are higher than the theoretical performance bound. This is surprising at the first glance but can well be explained: According to Section C of the *supplement document* [Krü09], SVQ in a sequential approach is superior to the parallel approach, and a combination with a gain SQ optimized for a Gaussian source naturally leads to a higher SNR for Gaussian

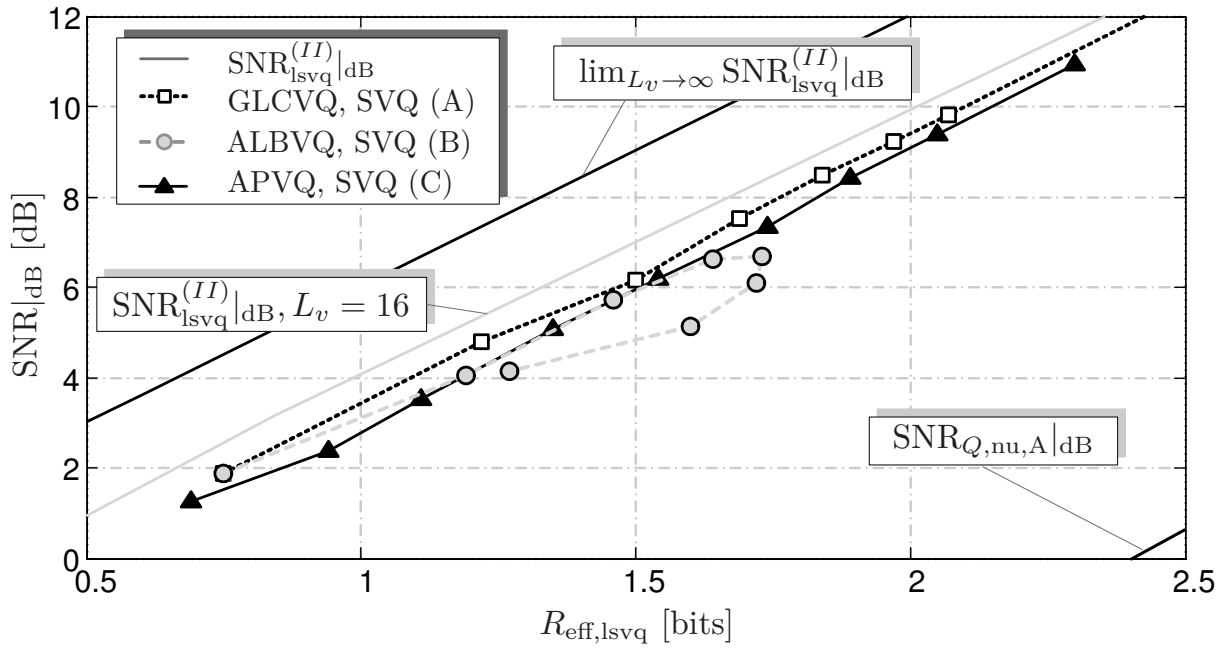


Figure 4.23: SNR for LSVQ over the effective bit rate for the GLCVQ, the ALBVQ, and the APVQ ($L_v = 16$). The curve for the ALBVQ is for $N_{(B)} = 2, 4, 6, 8, 10, 12, 14, 15$. The A-Law constant is set to $A = 5000$.

sources. As a conclusion of the comparison, despite the principal advantages of the Leech Lattice based approach, the GLCVQ in 24 dimensions reaches a performance which is asymptotically only 0.2 dB lower than the SNR for the Leech Lattice SVQ.

4.4.4.5 Visualization of the Measured Quantization Error

The distribution of the absolute value of the quantization error vector was computed in Section 4.3.4 and is a tool for the qualitative visualization of average quantization cell shapes. In Figure 4.25 a) and b), two example measured distributions of the absolute quantization error vectors are shown for $L_v = 8$. The example in Figure 4.25 a) was computed from the histogram of the absolute value of the quantization error for the APVQ and the example in Figure 4.25 b) from the histogram related to the GLCVQ. The theoretical result for the PDF of the absolute value of the quantization error derived in Section 4.3.4 for the "idealized" SVQ is shown as a reference and was computed for the same number of spherical codevectors, N_{svq}, with respect to Figure 4.25 a) and b), respectively.
From the comparison of the two plots, obviously in b), the distribution is more similar to the theoretical distribution than in a). This observation is well consistent with the measured SNR values in comparison to the SNR for the "idealized" SVQ as presented in Table 4.4.

4.5 Discussion

In this chapter, Logarithmic Spherical Vector Quantization (LSVQ) was investigated. In the first part, the principle of LSVQ was introduced, and theoretical

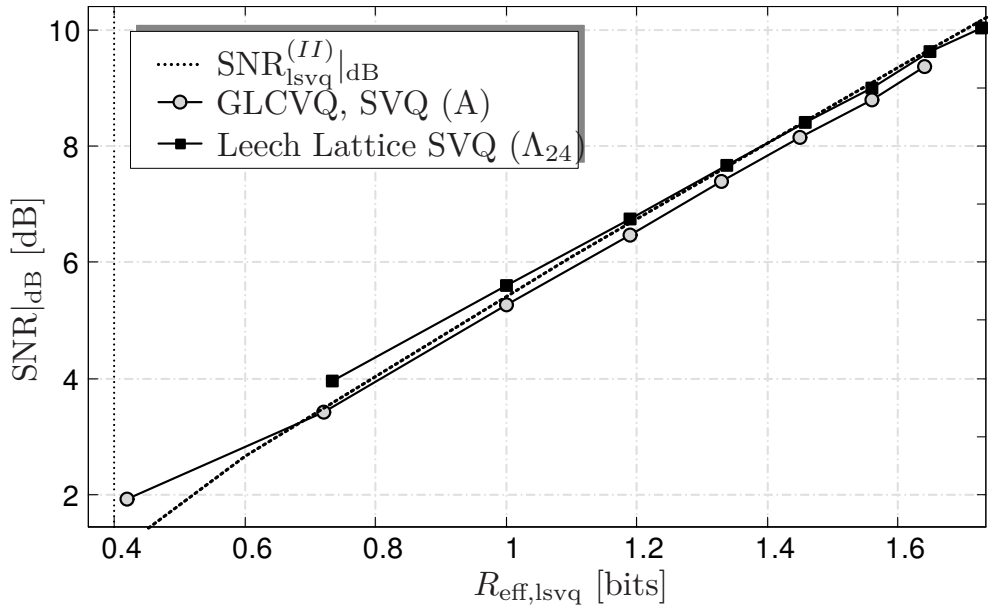
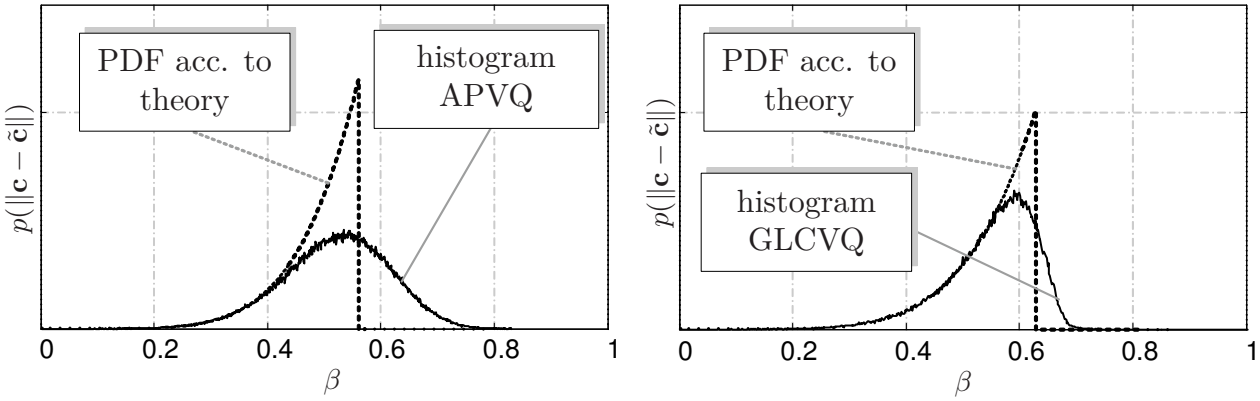


Figure 4.24: Comparison of the measured SNR related to the GLCVQ (Candidate SVQ (A)) in comparison to the results from [AB88] for a Leech lattice (Λ_{24}) based SVQ.

	Example SVQ (C) (SNR in dB)	Example SVQ (A) (SNR in dB)
SNR for "idealized" SVQ	6.26 dB	5.31 dB
measured SNR	5.64 dB	5.21 dB
effective bit rate $R_{\text{eff,svq}}$	1.12 bits	0.988 bits
number of spherical codevectors	502	240

Table 4.4: SNR for SVQ calculated from the measured and the theoretical PDF of the absolute value of the quantization error.

results were derived. At first it was shown that the SNR related to LSVQ is independent from the PDF of the signal to be quantized for the assumption of high bit rates. Following, an upper bound for the achievable SNR was derived based on the assumption of an "idealized" SVQ. Approximations introduced for high bit rates enabled to derive an analytical expression for the optimal allocation of bits to the gain and the shape component.
In the second part, the theoretical results from the first part were put in a context with practical realizations of LSVQ based on three different principles, the GLCVQ, the ALBVQ, and the APVQ introduced as candidates SVQ (A), (B), and (C), respectively. SNR curves were measured for uncorrelated Gaussian sources to assess all three candidates. As a conclusion of the measurements, it was shown that SVQ based on the algebraic codebooks nowadays used in speech coding, introduced



a) Histogram and PDF, Candidate SVQ (C)

b) Histogram and PDF, Candidate SVQ (A)

Figure 4.25: (Qualitative) measured histogram of the absolute value of the quantization error for two example spherical codes (APVQ and GLCVQ) and the PDF computed according to (4.61). The APVQ was operated at a bit rate of $R_{\text{eff,svq}} = 1.12$ bits, the GLCVQ at a bit rate of $R_{\text{eff,svq}} = 0.988$ bits for $L_v = 8$.

as the ALBVQ, candidate SVQ (B), has good performance for very low bit rates but can not be employed in every situation since the achievable bit rate is limited. The GLCVQ, candidate SVQ (A), has the highest quantization performance, and a nearest neighbor codevector search procedure can be realized with low computational complexity and memory consumption. It is also very flexible as it can be used for any vector dimension but outstanding performance is reached for dimensions which are multiples of eight. The design of the APVQ, candidate SVQ (C), is independent of the vector dimension which makes this the most flexible approach of all. The quantization performance is only marginally worse than that of the other candidates. The specific properties of the APVQ make it the best candidate for **efficient** combination with linear prediction techniques and is therefore also the basis for the SCELP and the W-SCELP codec.

5

Coding of Sources with Memory

In the previous chapter, it was shown that LSVQ is well suited for lossy source coding, and its performance is independent from the characteristics of the input signal in a wide dynamic range. However, being optimized for memoryless sources, LSVQ does not benefit from the correlation immanent to the input signal. In the literature, quantizers designed for memoryless sources are often combined with other techniques such as *Transform Coding* (TC) [HS63], [JN84], *Subband Coding* (SBC) [CR83], or *Linear Predictive Coding* (LPC) [MG76], [KP95] to exploit linear correlation.
TC and SBC are the basis for most state-of-the-art standardized audio codecs for music storage. Unfortunately, large transforms (in TC) and long finite impulse response (FIR) filters (in SBC) are in general required to achieve a sufficiently high spectral selectivity [PS00] so that these approaches are not suitable for audio coding with low algorithmic delay. Hence, LPC is the only technique which fulfills the delay constraints as it does not require any transform and, instead, is operated in the time domain on the basis of minimum phase digital filters. The use of LPC in **speech coding** is in general motivated by the fact that it mimics the physical process of human speech production [Fan60], [Fla72], [RS78].

In the first part of this chapter, LPC will be revisited, and the principle of linear prediction (LP) shall be combined with VQ in general and the LSVQ approach in particular. It will be motivated that LPC is indeed suitable also for low delay **audio coding**, but, in comparison to speech coding, new aspects have to be considered for low bit rates due to specific characteristics of audio signals and the adaptations to achieve low algorithmic delay. In this context, investigations are based on a novel *quantization noise production and propagation model* in the second part of this chapter and yield new theoretical results for closed-loop linear predictive quantization which, in contrast to the common theory on LPC for high bit rates, are also valid for lower bit rates. Resulting from the new model, a modified optimization

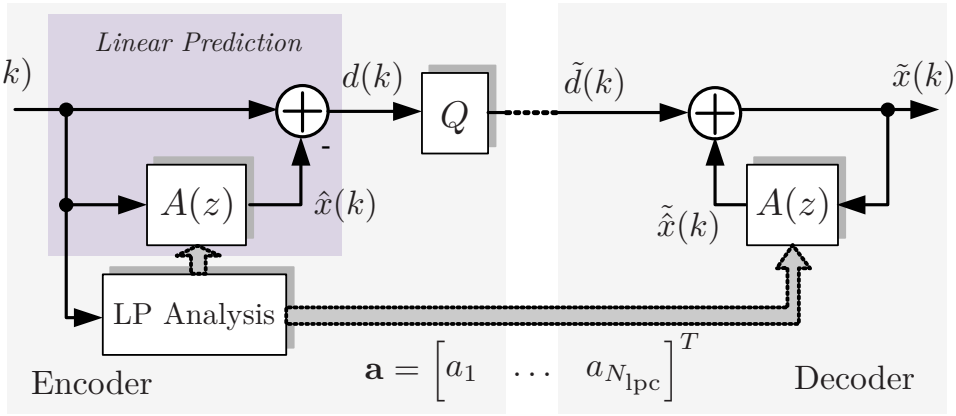


Figure 5.1: Principle of Linear Predictive Coding.

criterion for the computation of the involved filter coefficients is derived to realize the *reverse waterfilling* known from Section 2.3.

5.1 Linear Predictive Coding (LPC)

A typical scenario for Linear Predictive Coding (LPC) is shown in Figure 5.1.

5.1.1 Linear Prediction (LP)

Linear prediction (LP) [AS67], [MG76], [VM06] is a functional component of LPC, shown in Figure 5.1 as the gray block in the encoder. The target of linear prediction is to decorrelate the input signal $x(k)$ by means of linear filtering. For this purpose an estimate $\hat{x}(k)$ is calculated on the basis of previous signal values of $x(k)$ in the encoder as

$$\hat{x}(k) = \sum_{i=1}^{N_{\text{lpc}}} a_i \cdot x(k-i) \tag{5.1}$$

with the N_{lpc} LP coefficients a_i and the estimation block denoted by the system function $A(z)$ in Figure 5.1. The signal $d(k)$ is the estimation error or *LP residual signal* and hence the output of the *LP analysis filter* with system function

$$H_A(z) = 1 - A(z) = 1 - \sum_{i=1}^{N_{\text{lpc}}} a_i \cdot z^{-i}. \tag{5.2}$$

The filter coefficients a_i are calculated in the *LP analysis* block with the optimization target to minimize the variance (zero mean and stationarity of the signal $x(k)$ are assumed) of the LP residual signal $d(k)$,

$$E\{d^2(k)\} = E\{(x(k) - \hat{x}(k))^2\} \rightarrow \min. \tag{5.3}$$

The optimal coefficients a_i are determined as the solution of the *normal* linear equations,

$$\begin{bmatrix} \varphi_{x,x}(1) \\ \dots \\ \varphi_{x,x}(N_{\text{LPC}}) \end{bmatrix} = \begin{bmatrix} \varphi_{x,x}(0) & \dots & \varphi_{x,x}(1-N_{\text{LPC}}) \\ \varphi_{x,x}(1) & \dots & \varphi_{x,x}(2-N_{\text{LPC}}) \\ \dots & \dots & \dots \\ \varphi_{x,x}(N_{\text{LPC}}-1) & \dots & \varphi_{x,x}(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_{N_{\text{LPC}}} \end{bmatrix} \tag{5.4}$$

with

$$\varphi_{x,x}(\kappa) = E\{x(k) \cdot x(k+\kappa)\} \tag{5.5}$$

as the autocorrelation function (ACF) related to $x(k)$. The autocorrelation matrix has symmetric Toeplitz form, and the normalized linear equations can be efficiently solved by means of the Levinson-Durbin algorithm [Lev47][Dur60] to produce the optimal LP coefficients.

5.1.2 Block Adaptive LP

Since practical audio and speech signals are only short-term stationary, the LP coefficients are computed for short segments of the input signal in most modern speech codecs. In this context, the autocorrelation function is approximated by the short-term (energy) autocorrelation coefficients, denoted as the *autocorrelation method* [Mak75],

$$\hat{\varphi}_{x,x}(\kappa) \approx \varphi_{x,x}(\kappa) \tag{5.6}$$

with

$$\hat{\varphi}_{x,x}(\kappa) = \sum_{i=0}^{L_{\text{lpc}}-\kappa-1} x_w(i-\kappa) \cdot x_w(i). \tag{5.7}$$

$x_w(i)$ is an overlapping signal segment of length L_{lpc} which is extracted from the input signal $x(k)$ and often weighted by a window function $w(k)$ [KP95].
Alternative approaches for the calculation of the coefficients a_i are the *auto covariance method*, the *Burg algorithm* [dWB00] and also backward adaptive approaches, e.g., based on gradient descend algorithms [VM06]. Also, different realizations of the LP analysis filter in a direct form or a lattice based approach are known from the literature, e.g. [Kei06]. The alternative approaches shall not be discussed in the following.

5.1.3 LP and Quantization

In the LPC scheme in Figure 5.1, the LP residual signal $d(k)$ is quantized in the encoder. In order to reconstruct it at the decoder side, the quantized LP residual must be transformed into a binary index which is then transmitted to the decoder.

For the sake of simplicity, however, the transmission of the index is not shown in the figure.
The quantized LP residual signal in the decoder is signal

$$\tilde{d}(k) = \Delta(k) + d(k). \tag{5.8}$$

In that context, the quantizer is in general modeled as an additive noise source with the quantization error signal $\Delta(k)$. In the following, the quantization is at first based on SQ. This concept will then later be extended toward VQ in general in Section 5.1.7 and LSVQ in Section 5.1.8.
From the reconstructed LP residual signal $\tilde{d}(k)$, the decoded version of the input signal $\tilde{x}(k)$ is produced as the output of the *LP synthesis filter* with system function

$$H_S(z) = \frac{1}{1 - A(z)} = (H_A(z))^{-1}. \tag{5.9}$$

Due to the introduced quantization error, the *quantization SNR* is defined as

$$\mathrm{SNR}_0 = \frac{E\{d^2(k)\}}{E\{(d(k) - \tilde{d}(k))^2\}} \tag{5.10}$$

and characterizes the employed quantizer. Considering the impact of the quantizer with respect to the encoder input and the decoder output signals, SNR_0 must be distinguished from the *overall SNR*,

$$\mathrm{SNR} = \frac{E\{x^2(k)\}}{E\{(\tilde{x}(k) - x(k))^2\}} \tag{5.11}$$

related to the *processed quantization noise in the decoder output signal*. In this context the most important questions to be answered in the following are:

- What is the maximum achievable overall SNR given a specific quantizer and hence fixed quantization SNR_0?
- How do we achieve this maximum overall SNR?

In order to have access to the LP coefficients at the decoder side, in block adaptive LPC, the filter coefficients must be transmitted as side information. The transmission of these coefficients, denoted as vector $\mathbf{a} = \begin{bmatrix} a_1 & \dots & a_{N_{\mathrm{lpc}}} \end{bmatrix}^T$ in Figure 5.1, is assumed to be possible at low bit rates [PA93] and will not be discussed in the following. A theoretical analysis of the distribution of rate to the LP coefficients and the LP residual signal is given in [KO07]. A new practical concept for the efficient quantization of LP coefficients in the context of audio signals is described in [KSV06].

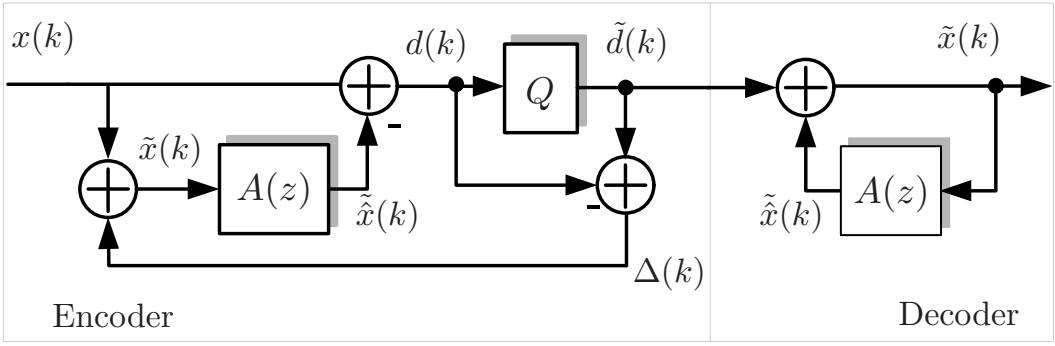
5.1.4 LPC and Audio Signals

The use of linear prediction in speech coding is often motivated by the modeling of the physics of human speech signal generation as an auto-regressive (AR) process: The excitation source of this process corresponds to the sound wave produced by the airflow through the vocal folds, and the evolution of formant frequencies in the human vocal tract is modeled by a time variant AR (all-pole) filter [MG76][VM06]. In contrast to this, no generic approach is known to model the variety of audio signals. Nevertheless, there are good arguments that linear prediction is also a promising candidate for audio coding: At first, a lot of natural and artificial audio signals are produced in a way very similar to the principles of human speech generation, e.g., in the context of a source-filter model as known from music synthesizers [Moo65], [FR00]. Inaccuracies of the AR model are also known from speech coding, e.g., the disregard of the nasal cavity, and do not lead to significant problems there either. Secondly, an audio signal is in general characterized by its local spectral maxima. If those are very strong, the rest of the signal may even be inaudible for humans [Zwi82]. In that context, the role of linear prediction is to efficiently represent the spectral envelope of a signal and especially the local spectral maxima by an all-pole filter rather than the identification of speech model parameters. In the context of Figure 5.1, the approximation of the spectral envelope of the input signal is realized by the LP synthesis filter $H_S(z)$ in the decoder block. Accounting for its role to approximate the spectral envelope of the input signal, $H_S(\Omega)$ computed from $H_S(z)$ for $z = e^{j\Omega}$ will be denoted also as the *LP filter spectrum* in the following.
The length of the LP analysis signal segments has a strong impact on the overall algorithmic delay of the codec and is therefore usually short to achieve a sufficiently low algorithmic in low delay audio coding, e.g., $L_{\text{lpc}} \approxeq 10$ ms. In comparison to this, in speech coding, a higher algorithmic delay is allowed so that the analysis segment lengths are in general longer, e.g., $L_{\text{lpc}} \approxeq 20 - 25$ ms.

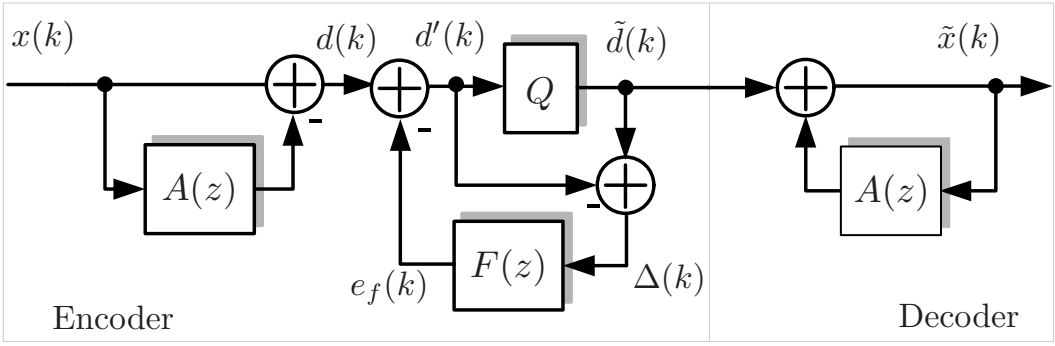
5.1.5 Zero-Mean Property of the LP Filter Spectrum

A property of the linear prediction analysis and synthesis filters which will be important in the following is denoted as the *zero-mean property*: Given the LP filter spectrum as the magnitude spectrum $|H_s(\Omega)|$ related to the system function $H_s(z)$, the average of the logarithmic magnitude spectrum is zero if the poles of $H_s(z)$ are located inside the unit circle [MG76],

$$\int_{-\pi}^{\pi} \ln(|H_s(\Omega)|)\frac{\mathrm{d}\Omega}{2\pi} = 0. \tag{5.12}$$



a) Closed-loop encoder and decoder



b) Generalized closed-loop encoder and decoder

Figure 5.2: Closed-loop LPC.

5.1.6 Closed-loop SQ

In Figure 5.1, in the encoder, the estimation signal $\hat{x}(k)$ is calculated from the input signal $x(k)$, whereas the estimation signal $\tilde{\hat{x}}(k)$ in the decoder is calculated from the quantized output signal $\tilde{x}(k)$ because $x(k)$ is not available in the decoder. In order to avoid this mismatch, the decoder can be simulated in the encoder to make the quantized output signal $\tilde{x}(k)$ available as the basis for the estimation signal in the encoder as well. In Figure 5.2 a), a modified realization of the LPC encoder is therefore shown, also known as *quantization in the loop* or *closed-loop quantization*. By feeding back the quantization error signal $\Delta(k)$, the estimated signal $\tilde{\hat{x}}(k)$ is now available also in the encoder. A modification of the decoder is not required. In contrast to this, the codec structure in Figure 5.1 is denoted as *open-loop quantization*.
In Figure 5.2 b), a generalization of a) is shown with the error weighting (or noise feedback) filter with system function $F(z)$. Assuming that the z-transforms related to all signals and also the quantization noise exist, the z-transform of the overall signal reconstruction error in the decoder output is

$$\tilde{x}(k) - x(k) \circ\!\!-\!\!\bullet\, \tilde{X}(z) - X(z) = \frac{1 - F(z)}{1 - A(z)} \cdot \Delta(z). \tag{5.13}$$

The quantization error signal $\Delta(k)$ is commonly assumed to be spectrally flat. Correspondingly, the spectral envelope of the processed version of the quantization

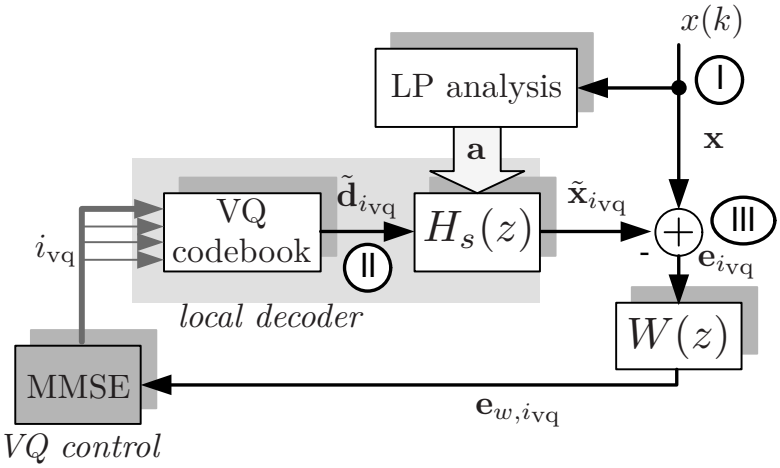


Figure 5.3: Principle of a Code-Excited Linear Prediction (CELP) Encoder.

noise in the reconstructed decoder output signal can be directly controlled by system function $F(z)$. This principle is known as *noise-shaping* (NS). A conventional technique known from the literature [SAH79] is to derive $F(z)$ from $A(z)$,

$$F(z) = F_{\mathrm{conv}}(z) = A(z/\gamma) \text{ with } 0 \leq \gamma \leq 1.0. \tag{5.14}$$

The constant γ controls the shift of the roots of the filter system function $1 - F(z)$ toward the origin in the z-domain compared to the roots of $1 - A(z)$ to create a flattened version of the LP filter spectrum, e.g., [VM06].
In the next section it will be shown that Figure 5.2 b) is equivalent to the *analysis-by-synthesis* approach in linear predictive VQ (LPVQ) under certain conditions and for the special case of vector dimension $L_v = 1$. The error weighting filter $F(z)$ will be discussed in detail in Sections 5.1.7.2 and 5.2.6.

5.1.7 Code-Excited Linear Prediction (CELP)

To combine linear prediction and VQ is straightforward in the case of **open-loop** quantization. The SQ (block Q in Figure 5.1) is simply replaced by the VQ, and the quantization is based on a nearest neighbor vector search. This approach, however, can not fully exploit linear dependencies as shown in [KV05].
Following the **closed-loop** approach, the aggregation of samples of the LP residual signal into vectors for quantization is contradictory to a sample-by-sample linear filtering of the quantization error in the feedback loop of Figure 5.2. Therefore, a new method for closed-loop linear predictive VQ was proposed in [SA85]. It is denoted as Code-Excited Linear Prediction (CELP), and forms the basis for today's most efficient speech coding algorithms, e.g., [ETS00], [ITU96]. A typical CELP encoder is shown in Figure 5.3. For the quantization of the signal $x(k)$, it is assumed that the LP coefficients have been determined previously in the LP analysis block (refer to Section 5.1.2). For the CELP encoding, all (non-overlapping) sequences of signals are transformed into vectors of dimension L_v. Note that it is a common

practice that one set of LP coefficients is computed for more than one signal vector, hence $L_v \neq L_{\text{lpc}}$ in general. In order to determine the optimal VQ codevector index $i_{Q,\text{vq}}$ for one signal vector $\mathbf{x}$, in the *local decoder* all N_{vq} codevectors $\tilde{\mathbf{d}}_{i_{\text{vq}}}$ taken from the *VQ codebook* are tested as candidate vectors for the LP residual signal. Each candidate vector is addressed by the index i_{vq} by the *VQ control* block and transformed into a signal candidate vector $\tilde{\mathbf{x}}_{i_{\text{vq}}}$ by means of filtering in the LP synthesis filter $H_S(z)$. The difference between $\mathbf{x}$ and $\tilde{\mathbf{x}}_{i_{\text{vq}}}$ (position III) is the error vector $\mathbf{e}_{i_{\text{vq}}}$. The weighted error vector $\mathbf{e}_{w,i_{\text{vq}}}$ is produced by filtering $\mathbf{e}_{i_{\text{vq}}}$ in the *error weighting filter* with system function $W(z)$ to be discussed later. In order to find the optimal codevector index, all weighted error vectors are compared to find that candidate which produces the (weighted) minimum mean squared error (MMSE),

$$i_{Q,\text{vq}} = \underset{0 \leq i_{\text{vq}} < N_{\text{vq}}}{\arg\min} \|\mathbf{e}_{w,i_{\text{vq}}}\|^2. \tag{5.15}$$

The generation of codevectors, the filtering of each candidate, the computation of the weighted error vectors, and the determination of the optimal codevector index are denoted as the *index iteration procedure* in the following. The overall computational complexity can be very high as the number of codevectors in the VQ codebook grows exponentially with the effective bit rate $R_{\text{eff,vq}}$ and the dimension L_v.

5.1.7.1 Modification of the CELP Approach

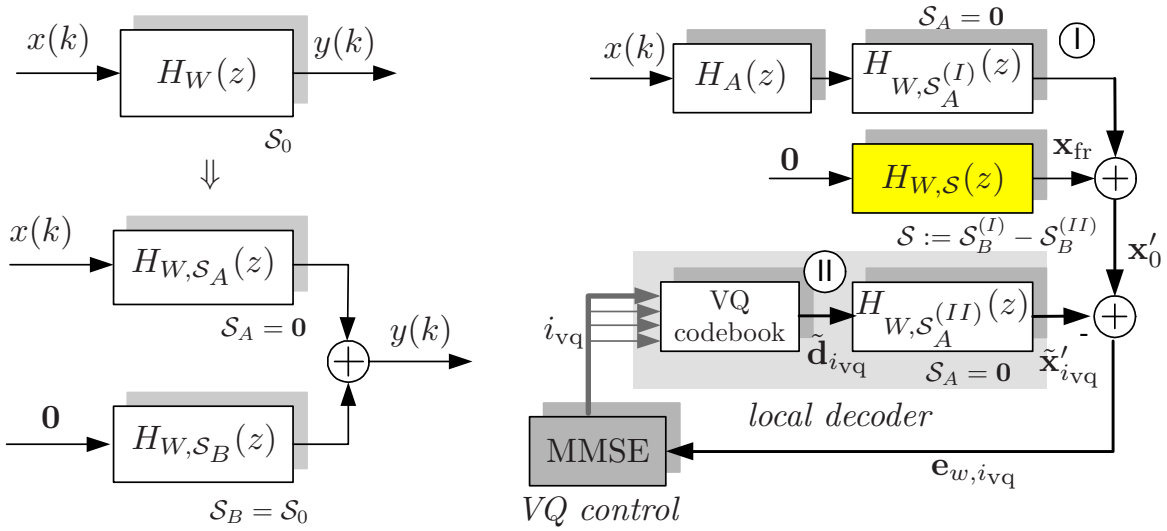
Targeting the reduction of the computational complexity involved in the CELP approach, at first, the encoder in Figure 5.3 is modified as proposed in [KV06b]. The result is shown in Figure 5.5 and will be explained in the following. Note that for the sake of simplicity, the LP analysis block is no longer shown in this figure. Considering the required modifications, at first the filter with system function

$$H_W(z) = H_S(z) \cdot W(z) \tag{5.16}$$

is introduced to combine the LP synthesis filter $H_S(z)$ and the weighting filter $W(z)$ in the *combined weighting filter* with system function $H_W(z)$. For the combination of both filters, in Figure 5.3, the weighting filter is relocated in each of the signal paths labeled as I and II in Figure 5.3. In path I, to form the combined weighting filter, the LP analysis filter $H_A(z)$ must be additionally introduced as shown in Figure 5.5 at position I since

$$W(z) = H_A(z) \cdot \underbrace{H_S(z) \cdot W(z)}_{H_W(z)} \qquad \text{with} \qquad H_A(z) \cdot H_S(z) = 1. \tag{5.17}$$

Due to the possibly long tails of the impulse responses of the combined weighting filter which can significantly exceed the vector dimension, the quantization of one vector has an impact on the quantization of successive signal vectors. In this context, it is beneficial to pry open the interaction across vector bounds so that the



a) Decomposition of Filters. b) Resulting CELP Encoder.

Figure 5.4: Decomposition of the combined weighting filter $H_W(z)$ with respect to the filter states (a)) and resulting modified CELP encoder (b)).

"history" of previously quantized vectors can be handled independently from the actual index iteration procedure. Consequently, the signal related to the "history" is computed only once for all codevector candidates rather than for all candidates. For this reason, the combined weighting filters $H_W(z)$ in the signal paths I and II are decomposed into two filters $H_{W,\mathcal{S}_A}(z)$ and $H_{W,\mathcal{S}_B}(z)$ as illustrated by Figure 5.4 a).

- Filter $H_{W,\mathcal{S}_A^{(I)}}(z)$ and $H_{W,\mathcal{S}_A^{(II)}}(z)$ hold states which are zero: $\mathcal{S}_A^{(I)} = \mathbf{0}$ and $\mathcal{S}_A^{(II)} = \mathbf{0}$, respectively.

- The filter $H_{W,\mathcal{S}_B^{(I)}}(z)$ and $H_{W,\mathcal{S}_B^{(II)}}(z)$ hold the states $\mathcal{S}_0^{(I)}$ and $\mathcal{S}_0^{(II)}$ from $H_W(z)$ in signal path I and II at the beginning of the quantization procedure for one vector $\mathbf{x}$. Both are fed by a vector composed of L_v zeros. $H^{(I)}_{W,\mathcal{S}_B}(z)$ in path I and $H^{(II)}_{W,\mathcal{S}_B}(z)$ in path II are linear and therefore can be realized as one filter $H_{W,\mathcal{S}}(z)$ with the superposed states

$$\mathcal{S} := \mathcal{S}_B^{(I)} - \mathcal{S}_B^{(II)}. \tag{5.18}$$

The resulting modified version of the CELP encoder is shown in Figure 5.4 b). The block $H_{W,\mathcal{S}}(z)$ with the combined states $\mathcal{S}$ from (5.18) is highlighted by the yellow color. The input signal of this block is a vector composed of zeros, and the states depend on the "history" of previously quantized vectors only (and not on the quantization process of the current vector). $H_{W,\mathcal{S}}(z)$ is handled outside the actual index iteration procedure with the corresponding output denoted as the *filter*

ringing signal $\mathbf{x}_{\mathrm{fr}}$.
The remaining filters $H_{W,\mathcal{S}_A^{(I/II)}}(z)$ in path I and II with the states set to zero do not depend on any quantization history and therefore can be written as matrix

$$\mathbf{H}_W = \begin{bmatrix} h_W(0) & 0 & \dots & 0 \\ h_W(1) & h_W(0) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ h_W(L_v-1) & h_W(L_v-2) & \dots & h_W(0) \end{bmatrix}. \tag{5.19}$$

The filtering is hence realized as a convolution of the input signal vectors with the truncated impulse response

$$h_W(k) \circ\!\!-\!\!\bullet\, H_W(z) \qquad \forall \qquad 0 \le k < L_v. \tag{5.20}$$

In the next step, the modifications are completed by substituting

$$\mathbf{d}_{\mathrm{fr}} = \mathbf{H}_W^{-1} \cdot \mathbf{x}_{\mathrm{fr}} \tag{5.21}$$

and setting

$$\mathbf{d}'' = \mathbf{d} + \mathbf{d}_{\mathrm{fr}} = \mathbf{d} + \mathbf{H}_W^{-1} \cdot \mathbf{x}_{\mathrm{fr}}, \tag{5.22}$$

shown at position IV in Figure 5.5. The weighted error vector can be expressed as

$$\mathbf{e}_{w,i_{\mathrm{vq}}} = \mathbf{x}'_0 - \tilde{\mathbf{x}}'_{i_{\mathrm{vq}}} = \mathbf{x}'_0 - \mathbf{H}_W \cdot \tilde{\mathbf{d}}_{i_{\mathrm{vq}}} = \mathbf{H}_W \cdot (\mathbf{d}'' - \tilde{\mathbf{d}}_{i_{\mathrm{vq}}}). \tag{5.23}$$

The search procedure to determine the optimal index (5.15) can hence be expressed as

$$\begin{aligned} i_{Q,\mathrm{vq}} &= \underset{0 \le i_{\mathrm{vq}} < N_{\mathrm{vq}}}{\arg\min}\ (\mathbf{x}'_0 - \mathbf{H}_W \cdot \tilde{\mathbf{d}}_{i_{\mathrm{vq}}})^T \cdot (\mathbf{x}'_0 - \mathbf{H}_W \cdot \tilde{\mathbf{d}}_{i_{\mathrm{vq}}}) \\ &= \underset{0 \le i_{\mathrm{vq}} < N_{\mathrm{vq}}}{\arg\min}\ (\mathbf{d}'' - \tilde{\mathbf{d}}_{i_{\mathrm{vq}}})^T \cdot \mathbf{H}_W^T \cdot \mathbf{H}_W \cdot (\mathbf{d}'' - \tilde{\mathbf{d}}_{i_{\mathrm{vq}}}) \end{aligned} \tag{5.24}$$

Both matrices, $\mathbf{H}_W$ and $\mathbf{H}_W^{-1}$, and the signal vector $\mathbf{d}''$ can be easily computed prior to the index iteration procedure. Due to the weighting of the quantization error, in comparison to (3.30) and (4.13), (5.24) can no longer be realized as a *nearest neighbor* quantization.
Once the optimal codevector has been found, in order to properly prepare the quantization of the next vector, the states $\mathcal{S}$ must be updated by filtering the LP residual error vector

$$\mathbf{\Delta} = \mathbf{d} - \tilde{\mathbf{d}}_{i_{Q,\mathrm{vq}}}. \tag{5.25}$$

This state update procedure is shown as the *update* arrow in Figure 5.5. Note that in the figure, $\tilde{\mathbf{d}}_{i_{\mathrm{vq}}}$ is used to address all possible codebook entries in the index iteration procedure whereas $\tilde{\mathbf{d}}_{i_{Q,\mathrm{vq}}}$ denotes the optimal codevector which was selected.
Due to the modified structure of the CELP encoder, the codevector search procedure can be realized as a three step procedure involving the

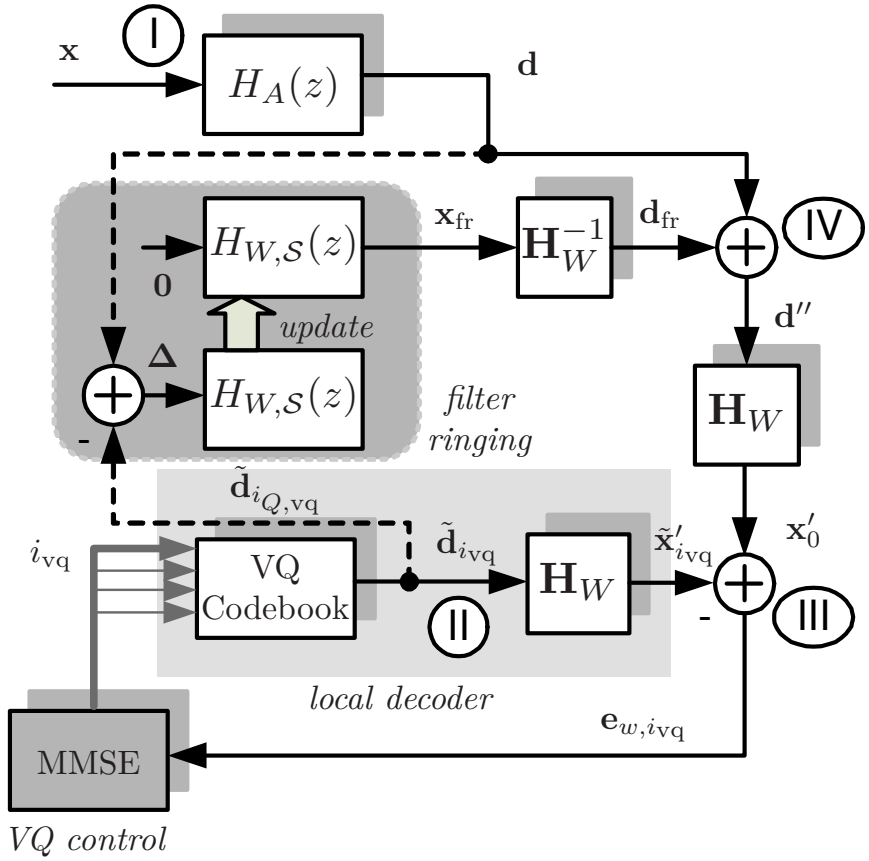


Figure 5.5: Modified approach for the Code-Excited Linear Prediction (CELP) encoder.

1. **Pre-processing**: Computation of the matrices $\mathbf{H}_W$ (5.19) and $\mathbf{H}_W^{-1}$ and production of the filter ringing signal $\mathbf{x}_{\mathrm{fr}}$ and its transformed version $\mathbf{d}_{\mathrm{fr}}$ in (5.21) to compute $\mathbf{d}''$.

2. **Weighted codevector search** (5.24) involving matrix $\mathbf{H}_W$

3. **Post-processing**: Filter update procedure based on $\boldsymbol{\Delta}$ from (5.25).

Due to the modified CELP encoder, the filtering of each codevector candidate is replaced by a convolution which leads to a reduction of the involved computational complexity compared to the approach as described in Section 5.1.7 for vector dimensions and orders of the combined weighting filter as used in the SCELP codec (Section 6). More important than this, however, is that the signal $\mathbf{d}''$ is available which is an important and necessary starting point for the efficient realizations of the index iteration procedure in the SCELP codec to be investigated in Section 6.1.3. In addition to that, the three step procedure enables to better understand methods for low complexity realizations of CELP coding as discussed in Section E of the *supplement document* [Krü09] and to enhance the overall coding concept by means of frequency warping which is described in Section 6.2.

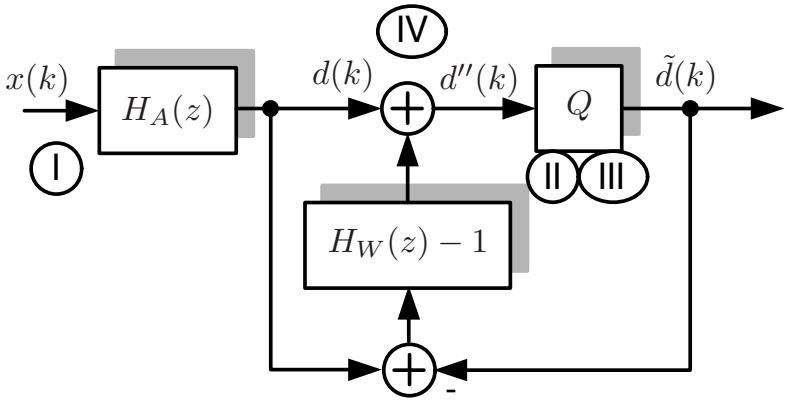


Figure 5.6: Modified CELP encoder for $L_v = 1$.

5.1.7.2 Error Weighting Filter $W(z)$

The error weighting filter $W(z)$ in CELP coding is the generalization of the error weighting filter $F(z)$ (5.14) for LPC and SQ to control the spectral envelope of the processed quantization noise in the decoder output signal. It is often chosen as

$$W(z) = \frac{1 - A(z/\gamma_1)}{1 - A(z/\gamma_2)} \tag{5.26}$$

with $0 < \gamma_2 \leq \gamma_1 \leq 1.0$. It is a common assumption that the signal $e_w(k)$, given in vector notation as $\mathbf{e}_{w,i_{\mathrm{VQ}}}$, is spectrally flat. In that case the spectral envelope related to the processed quantization noise in the decoder output signal is a function of the filter system function $W^{-1}(z)$. Putting this into a relation to (5.14), $W(z)$ should be chosen as

$$W(z) = \frac{1 - A(z)}{1 - F(z)} \tag{5.27}$$

with $\gamma_1 = 1.0$ and $\gamma_2 = \gamma$ in (5.26) to be equivalent to the SQ approach from Section 5.1.6.

5.1.7.3 Comparison of SQ based LPC and CELP Encoder for $L_v = 1$

For a vector dimension of $L_v = 1$ and choosing $W(z)$ according to (5.27), the CELP encoder in its modified structure from Figure 5.5 is identical to the generalized closed-loop encoder as shown in Figure 5.2 b). This can be shown as follows:
Given a vector dimension of $L_v = 1$, the matrices containing the truncated impulse response of the combined weighting filter are

$$\mathbf{H}_W = \mathbf{H}_W^{-1} = 1. \tag{5.28}$$

The filter ringing part in Figure 5.5, including the state update operation, can be written as a linear filter with system function $H_W(z) - 1$ as shown in Figure 5.6. For $L_v = 1$ all vectors are scalars and can be rewritten as a function of time index k. Since the matrices $\mathbf{H}_W$ and its inverse do not have any impact due to (5.28), the

index iteration procedure is identical to conventional (nearest neighbor) SQ (block Q at position II/III in Figure 5.6). Assuming that the z-transforms related to all signals exists, the resulting block diagram is analyzed in the z-domain: Setting the weighting filter according to (5.27), the combined weighting filter is

$$H_W(z) = \frac{1}{1-A(z)} \cdot \frac{1-A(z)}{1-F(z)} = \frac{1}{1-F(z)}. \tag{5.29}$$

The z-transform of signal $d''(k)$ is

$$\begin{aligned} d''(k) \circ\!\!-\!\!\bullet\, D''(z) &= D(z) + \left(\frac{1}{1-F(z)} - 1\right) \cdot (D(z) - \tilde{D}(z)) \\ D''(z) &= D(z) - F(z) \cdot (\tilde{D}(z) - D''(z)) \end{aligned} \tag{5.30}$$

By comparing this result to Figure 5.2 b), it is obvious that $d'(k) = d''(k)$ so that both realizations are identical for $L_v = 1$.

5.1.7.4 Mapping of CELP Coding to an SQ based LPC Model

In both, LPC combined with SQ and with VQ, the quantizer transforms every sequence of samples of the LP residual signal $d(k)$ into a sequence of quantized output samples $\tilde{d}(k)$. Given signal vectors $\mathbf{d}$, $\mathbf{d}''$, and $\tilde{\mathbf{d}}_{i_{Q,\mathrm{vq}}}$ in the quantization process of a CELP encoder in the structure of Figure 5.5, the quantization reconstruction levels $\tilde{d}(k)$ and the quantization error $\Delta(k)$ can be reproduced on a sample-by-sample (or coordinate-by-coordinate) basis. Based on this vector decomposition, every CELP quantization process of arbitrary dimension can be mapped to the special case of $L_v = 1$ as shown in Figure 5.6 and, due to the the proof given in the previous section, to the SQ based LPC as shown in Figure 5.2 b). Necessary condition that both, the CELP approach and the SQ based closed-loop LPC approach, produce the same overall output signal $\tilde{x}(k)$ is that the SQ outputs the same quantization reconstruction levels $\tilde{d}(k)$ as the VQ and that the noise-shaping filter $F(z)$ is chosen according to (5.27). Therefore, a scalar model according to Figure 5.5 will be the common basis for all theoretical investigations on combined LP and SQ and CELP coding as described in the following.

But what is the advantage of the CELP approach compared to a closed-loop SQ approach in LPC then? At first, a VQ can be more flexible than a SQ as it does not rely on the same output alphabet for each vector coordinate. This was the basis for the shape advantage described in Section 3.2.1.2. Hence, in a practical application, the SQ can not necessarily output the same sequence $\tilde{d}(k)$ as the VQ. In the following, however, this practical aspect does not have any impact on the theoretical analysis of LPC[1].

In addition to that, an SQ determines each quantization reconstruction level one after the other according to the sequence of signal amplitudes to be quantized.

[1] If $\tilde{d}(k)$ is known, the SQ can always be assumed to be realized as a SQ with a "switchable scalar reconstruction alphabet". For example, the APVQ from Section 4.4.3 is based on this concept.

Therefore, if an optimal reconstruction level as the output at one moment proves itself as a bad choice later on, this first decision can never be withdrawn. In comparison to this, a VQ determines the optimal quantization reconstruction levels for all vector coordinates at the same time. In the context of the combination of VQ and LPC and the scalar LPC model, this principle enables to introduce an "implicit error weighting filter" which will be in detail discussed in Section 5.2.7. One step towards the correction of previously made decisions in SQ is the so-called delayed-decision quantization [JN84] which, however, shall not be discussed here.

5.1.8 CELP Coding and Gain-Shape Decomposition

So far, the modified CELP approach was developed for VQ in general. Overall goal, however, is to combine LPC with the concepts for LSVQ from Section 4.4. In LSVQ, the codevectors to approximate the LP residual signal $d(k)$ are denoted as $\tilde{\mathbf{d}}_{i_{\mathrm{lsvq}}}$. Each vector $\tilde{\mathbf{d}}_{i_{\mathrm{lsvq}}}$ is computed in analogy to (4.9) from a spherical codevector $\tilde{\mathbf{c}}_{i_{\mathrm{svq}}}$ and a quantizer reconstruction level for the radius, $\tilde{g}_{i_g}$,

$$\tilde{\mathbf{d}}_{i_{\mathrm{lsvq}}} = \tilde{g}_{i_g} \cdot \tilde{\mathbf{c}}_{i_{\mathrm{svq}}} \tag{5.31}$$

with i_{lsvq} to be calculated from i_g and i_{svq} according to (4.16). In analogy to the definition of LSVQ from Section 4.2.2, the decomposition of codevectors into a gain factor and a spherical codevector enables to realize an efficient codevector search procedure. Besides the parallel approach discussed in Section 4.2.2, the *joint* and the *sequential* approach as variants of LSVQ are described for a nearest neighbor quantization in Section C of the *supplement document* [Krü09]. Both approaches must be modified in the context of the **weighted codevector search procedure** given in (5.24) and will be briefly described in the following. More detailed investigations on these approaches and other aspects related to efficient codevector search procedures will be presented in the context of the SCELP codec in Section 6.1.3 and in Section E of the *supplement document* [Krü09].

5.1.8.1 The Joint Approach

The joint approach is used in most standardized CELP speech codecs and is described in, e.g., [Pau96], [HSW01]. The weighted codevector search procedure in (5.24) to find the optimal codevector index can be written as

$$i_{Q,\mathrm{svq}} = \underset{0 \le i_{\mathrm{svq}} < N_{\mathrm{svq}}}{\arg\min} \quad \mathcal{M}^{\mathrm{J}}_{i_{\mathrm{svq}}}. \tag{5.32}$$

with the metric $\mathcal{M}^{\mathrm{J}}_{i_{\mathrm{svq}}}$ given as

$$\begin{aligned} \mathcal{M}^{\mathrm{J}}_{i_{\mathrm{svq}}} &= (\mathbf{x}'_0 - \mathbf{H}_W \cdot g_{\mathrm{opt}} \cdot \tilde{\mathbf{c}}_{i_{\mathrm{svq}}})^T \cdot (\mathbf{x}'_0 - \mathbf{H}_W \cdot g_{\mathrm{opt}} \cdot \tilde{\mathbf{c}}_{i_{\mathrm{svq}}}) \\ &= \mathbf{x}_0'^{\,T} \mathbf{x}'_0 - 2 \cdot g_{\mathrm{opt}} \cdot \mathbf{x}_0'^{\,T} \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\mathrm{svq}}} + g^2_{\mathrm{opt}} \cdot \tilde{\mathbf{c}}^T_{i_{\mathrm{svq}}} \cdot \mathbf{H}^T_W \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\mathrm{svq}}}. \end{aligned} \tag{5.33}$$

In this equation, the optimal gain g_{opt} rather than the quantized version $\tilde{g}_{i_g}$ is used at first. g_{opt} is computed by setting the derivative of (5.33) with respect to g_{opt} to zero to yield

$$g_{\text{opt}} = \frac{\mathbf{x}_0'^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}}}{\tilde{\mathbf{c}}_{i_{\text{svq}}}^T \cdot \mathbf{H}_W^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}}}. \tag{5.34}$$

It is the optimal gain factor given a vector $\mathbf{x}_0'$ and a codevector $\tilde{\mathbf{c}}_{i_{\text{svq}}}$. Substituting g_{opt} in (5.33) yields

$$\mathcal{M}_{i_{\text{svq}}}^{\text{J}} = \mathbf{x}_0'^T \cdot \mathbf{x}_0' - \frac{(\mathbf{x}_0'^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}})^2}{\tilde{\mathbf{c}}_{i_{\text{svq}}}^T \cdot \mathbf{H}_W^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}}}. \tag{5.35}$$

Therefore, $\mathcal{M}_{i_{\text{svq}}}^{\text{J}}$ depends only on the spherical codevector $\tilde{\mathbf{c}}_{i_{\text{svq}}}$.
Since $\mathcal{M}_{i_{\text{svq}}}^{\text{J}}$ and $\mathbf{x}_0'^T \cdot \mathbf{x}_0'$ by definition are positive, the determination of the optimal spherical codevector can be written as

$$i_{Q,\text{svq}} = \underset{0 \le i_{\text{svq}} < N_{\text{svq}}}{\arg\max} \quad \mathcal{M}_{i_{\text{svq}}}^{\text{J}'} \tag{5.36}$$

involving the alternative metric

$$\mathcal{M}_{i_{\text{svq}}}^{\text{J}'} = \frac{(\mathbf{x}_0'^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}})^2}{\tilde{\mathbf{c}}_{i_{\text{svq}}}^T \cdot \mathbf{H}_W^T \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}}} \tag{5.37}$$

If the index $i_{Q,\text{svq}}$ and correspondingly the optimal spherical codevector $\tilde{\mathbf{c}}_{i_{Q,\text{svq}}}$ have been determined, the optimal gain factor is computed according to (5.34) and quantized in the A-Law SQ to produce the index $i_{Q,g}$.

5.1.8.2 The Sequential Approach

Given the signal $\mathbf{d}''$, in the first step of the sequential quantization procedure, a gain factor is computed as

$$g_{\mathbf{d}''} = \| \mathbf{d}'' \| \tag{5.38}$$

and quantized in the A-Law SQ to produce $\tilde{g}_{i_{Q,g}}$. In the next step, the quantized gain is used to find the optimal spherical codevector index as

$$i_{Q,\text{svq}} = \underset{0 \le i_{\text{svq}} < N_{\text{svq}}}{\arg\min} \quad \mathcal{M}_{i_{\text{svq}}}^{\text{S}} \tag{5.39}$$

based on the metric

$$\mathcal{M}_{i_{\text{svq}}}^{\text{S}} = (\mathbf{x}_0' - \tilde{g}_{i_{Q,g}} \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}})^T \cdot (\mathbf{x}_0' - \tilde{g}_{i_{Q,g}} \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{\text{svq}}}). \tag{5.40}$$

If the index $i_{Q,\mathrm{svq}}$ and correspondingly the optimal spherical codevector $\tilde{\mathbf{c}}_{i_{Q,\mathrm{svq}}}$ have been determined, finally, the optimal gain factor is computed in analogy to (5.34) as

$$g'_{\mathbf{d}''} = \frac{\mathbf{x}'_0 \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{i_{Q,\mathrm{svq}}}}{\tilde{\mathbf{c}}^T_{Q,i_{\mathrm{svq}}} \cdot \mathbf{H}^T_W \cdot \mathbf{H}_W \cdot \tilde{\mathbf{c}}_{Q,i_{\mathrm{svq}}}} \tag{5.41}$$

which is quantized to produce an update of the index $i_{Q,g}$. The second quantization of the gain factor is denoted as the "requantization".
It will be explained in Section 6 that the sequential approach is the more efficient technique for high performance quantization with low computational complexity in the SCELP codec (for more details also refer to Section E.1 of the *supplement document* [Krü09].).

5.2 Theoretical Analysis of LPC

Theoretical analyses of LPC for speech coding are in the most cases based on the assumption of high bit rates, e.g., [GR92]. In practical applications, the high rate theory is not sufficient to explain specific phenomena observed for low bit rates such as encoder instabilities. The observation of instabilities in LPC indeed seems to be very astonishing since for simple variants of LPC such as Delta Modulation [Ger72], Differential Pulse Code Modulation (DPCM), and Adaptive Differential Pulse Code Modulation (ADPCM) [GL77], [Kie82], [KD83] deterministic and stochastic stability has been proven. Despite these proofs, for the more complex approaches based on block adaptive LPC with higher order linear prediction and feedback of the quantization noise, unstable behavior at very low bit rates was already described in [Ata82]. As a solution there, a technique to manipulate the LP spectrum at high frequencies is proposed. In [JN84], it is described that in noise-feedback coding (NFC), which is identical to the closed-loop approach in Figure 5.2 b), a limiter is required to guarantee stability. It is stated that the codec tends to become unstable because of overload effects in the quantizer.
In order to explain the described phenomena, a new theoretical analysis of LPC is developed in the following. Key element of this analysis is a novel scalar *quantization noise production and propagation model* which is valid for open- and closed-loop SQ and VQ (CELP). In the new model, the error introduced by the quantization of the LP residual signal is modeled by a *gain-controlled additive noise source*, motivated by the results from Section 2 (rate distortion theory) and 4 (LSVQ) that optimal and universal quantizers produce a constant quantization SNR. The processing of the quantization noise by the combination of error weighting filter $F(z)/W(z)$ and LP synthesis filter $H_S(z)$ is modeled by a *noise propagation network*. The new model confirms the results known from the literature derived for high bit rates but is valid also for lower bit rates as it also accounts for the interaction between the feedback of the quantization error and the quantizer. This interaction is commonly

neglected in conventional analyses of LPC.
The model explains why closed-loop quantization can become unstable and enables to compute a theoretical overall coding SNR (5.11) which deviates from the SNR predicted by the conventional theory. As a conclusion, a modified concept for closed-loop quantization at low bit rates is finally derived based on a new optimization criterion which is well applicable in practice. It is shown that the described aspects are related to the effect of *reverse waterfilling* which was introduced in the context of the rate distortion theory in Section 2 but, so far, has not been intensively investigated in the context of LPC.

5.2.1 Prerequisites

The new scalar quantization noise production and propagation model assumes that all signals are stationary with zero mean in the following. The goal is to compute relations between signal variances to determine the overall SNR (5.11) as a function of the quantization SNR, SNR_0 (5.10).
The impact of all filters will be considered as *filtering gains*. The filtering gains are derived from the Wiener-Lee Relation and the Parseval Theorem [Lük95] which state that, if an uncorrelated stationary signal $x(k)$ with constant power spectral density (PSD)

$$\phi_x(\Omega) = \sigma_x^2 \tag{5.42}$$

is filtered by a filter with system function $H(z)$, the energy of the filter output signal $y(k)$ is

$$E\{y^2(k)\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_y(\Omega)\mathrm{d}\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\Omega)|^2\sigma_x^2\mathrm{d}\Omega. \tag{5.43}$$

In this context, the filtering gain $G_{x,y}$ is defined as the relation between the variances of the filter output and the filter input,

$$G_{x,y} := \frac{E\{y^2(k)\}}{E\{x^2(k)\}} = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(\Omega)|^2\mathrm{d}\Omega \tag{5.44}$$

5.2.2 Definition of the *Quantization Noise Production and Propagation Model*

The (quantization) noise (production and) propagation model is depicted in Figure 5.7 and follows the generalized closed-loop approach from Figure 5.2. In analogy to (5.11), the overall coding SNR is defined as the relation between the variance of the signal to be encoded $x(k)$ (position P_I in the figure) and that of the quantization noise in the decoder output, $(\tilde{x}(k) - x(k))$ (position P_{II} in the figure),

$$\mathrm{SNR}_{\mathrm{lpc}} = \frac{E\{x^2(k)\}}{E\{(\tilde{x}(k) - x(k))^2\}}. \tag{5.45}$$

The model consists of the following five components:

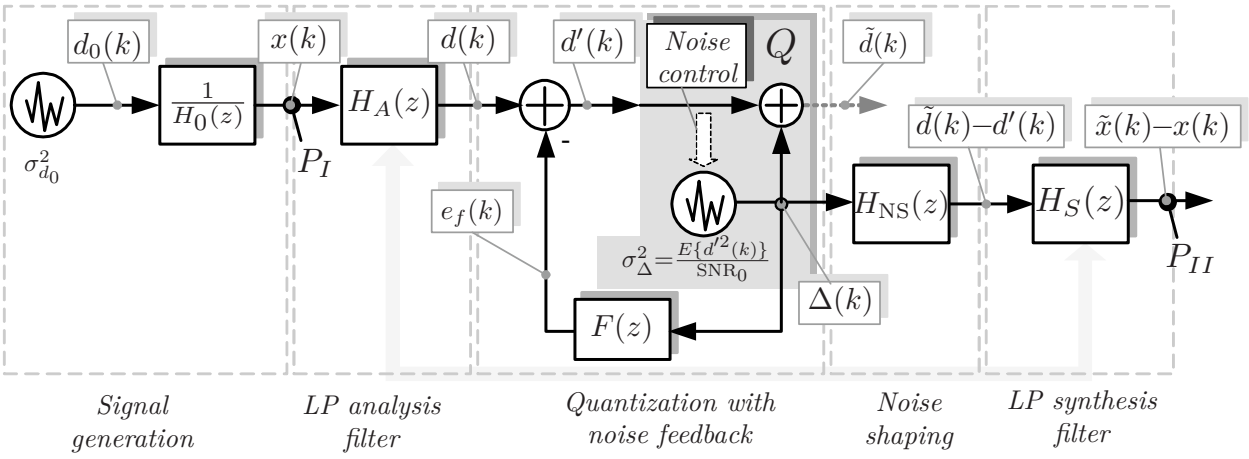


Figure 5.7: Noise production and propagation model for LPC. Only the propagation of the quantization noise in the decoder is considered rather than the reconstruction of the decoder output signal $\tilde{x}(k)$.

- **The signal generation block**: The stationary signal $x(k)$ to be coded is assumed to be the output of an AR process in the *signal generation* block. The AR process is realized as an AR (all-pole) filter of order N_{ar} with system function

$$H_{\mathrm{ar}}(z) = \frac{1}{H_0(z)} = \frac{1}{1 - \sum_{i=1}^{N_{\mathrm{ar}}} a_{\mathrm{ar},i} \cdot z^{-i}} \tag{5.46}$$

  with the AR coefficients $\mathbf{a}_{\mathrm{ar}} = \begin{bmatrix} a_{\mathrm{ar},1} & \cdots & a_{\mathrm{ar},N_{\mathrm{ar}}} \end{bmatrix}$ which is fed by an (uncorrelated) excitation signal $d_0(k)$ with variance $\sigma^2_{d_0}$ (and zero mean)[2]. Note that since all poles of $H_{\mathrm{ar}}(z)$ are located inside the unit circle, the corresponding magnitude spectrum has zero-mean property (refer to Section 5.1.5).

- **The LP analysis filter block**: The LP filter coefficients are computed from signal $x(k)$ in the LP analysis as described in Section 5.1.2 (which is not part of the figure) and used in the LP analysis filter $H_A(z)$ of order N_{lpc}. The output of the LP analysis filter $H_A(z)$ is the LP residual signal $d(k)$. The LP analysis filter is assumed to be of a similar order as $H_0(z)$ ($N_{\mathrm{lpc}} \approx N_{\mathrm{ar}}$). Then, $H_A(z)$ is a good approximation of $H_0(z)$:

$$H_A(z) \approx H_0(z). \tag{5.47}$$

  Therefore, signal $d(k)$ is similar to signal $d_0(k)$.

- **The quantization with noise feedback block**: This block consists of the quantizer Q and the error weighting filter $F(z)$. The signal $d'(k)$ to be quantized is computed from the signals $d(k)$, the output of the LP analysis filter, and signal $e_f(k)$. The quantizer output signal would be signal $\tilde{d}(k)$. However,

[2] A motivation why the results derived in the context of an AR model are also valid for the coding of audio signals was given in Section 5.1.4.

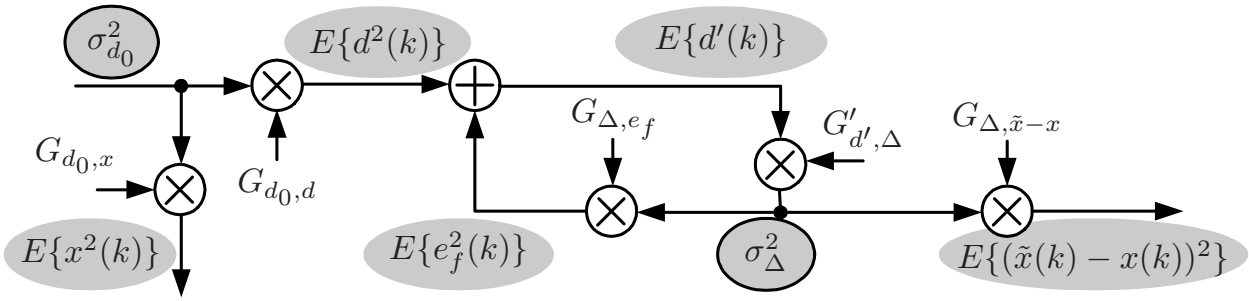


Figure 5.8: Block diagram based on variances and filtering gains. Variances marked by a black frame are related to spectrally flat (uncorrelated) signals. The variances related to signal $e_f(k)$ and signal $d(k)$ are assumed to be independent and therefore must be added.

signal $\tilde{d}(k)$ is not of importance in our noise propagation model and hence not considered in all subsequent blocks. Instead, only the quantization error signal $\Delta(k)$ is shown in Figure 5.7. In general, the quantization error is assumed to be independent of the input signal of the quantizer and uncorrelated if the bit rate is high enough, e.g. [JN84]. This is assumed to be true also for our model and will be discussed more in detail in Section 5.2.5. With respect to the fact that our quantizer produces a constant SNR_0, however, the **variance** of the quantization error depends on the **variance** of the quantizer input signal $d'(k)$. Therefore, the core quantizer is modeled as a power controlled noise source with variance $\sigma^2_\Delta = E\{d'^2(k)\}/\mathrm{SNR}_0$. The dependence between the variances of signal $d'(k)$ and $\Delta(k)$ is highlighted by the *Noise control* arrow in the figure. $\Delta(k)$ is filtered in the error weighting (noise feedback) filter $F(z)$ to produce signal $e_f(k)$.

- **The noise shaping block**: Signal $\tilde{d}(k)-d'(k)$ is a filtered version of the quantization noise $\Delta(k)$ in our noise propagation model due to the error weighting filter $F(z)$ in the encoder. With respect to Section 5.1.6, the involved transfer function to compute $\tilde{d}(k) - d'(k)$ from $\Delta(k)$ is $H_{\mathrm{NS}}(z) = 1 - F(z)$.

- **The LP synthesis filter**: The LP synthesis filter is the inverse of the LP analysis filter, $H_S(z) = (H_A(z))^{-1}$. In our model, it is assumed that the same coefficients are available in encoder and decoder and hence for the LP analysis and the synthesis filters.

For the computation of the overall SNR (5.45), Figure 5.7 is transformed into the diagram in Figure 5.8. In that diagram, instead of the signals and filters, the variances and filtering gains are shown. The signals $d_0(k)$ and $\Delta(k)$ are assumed to be uncorrelated signals and spectrally flat. Therefore the corresponding variances are highlighted by the circles with a black frame in the figure. Based on this

assumption, the filtering gains shown in the figure are calculated as follows:

$$G_{d_0,x} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1}{H_0(\Omega)}\right|^2 \mathrm{d}\Omega \tag{5.48}$$

$$G_{d_0,d} = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{H_A(\Omega)}{H_0(\Omega)}\right|^2 \mathrm{d}\Omega \tag{5.49}$$

$$G_{\Delta,e_f} = \frac{1}{2\pi}\int_{-\pi}^{\pi}|F(\Omega)|^2 \,\mathrm{d}\Omega \tag{5.50}$$

$$G_{\Delta,\tilde{x}-x} = \frac{1}{2\pi}\int_{-\pi}^{\pi}|H_{\mathrm{NS}}(\Omega)\cdot H_S(\Omega)|^2 \,\mathrm{d}\Omega = \frac{1}{2\pi}\int_{-\pi}^{\pi}\left|\frac{1-F(\Omega)}{1-A(\Omega)}\right|^2 \mathrm{d}\Omega \tag{5.51}$$

$$G'_{d',\Delta} = \frac{1}{\mathrm{SNR}_0} \tag{5.52}$$

Note that $H_{\mathrm{NS}}(z)$ and $H_S(z)$ are combined in a single filtering gain $G_{\Delta,\tilde{x}-x}$. G_{Δ,e_f} is the filtering gain with respect to the error weighting (noise feedback) filter $F(z)$ and will in the following be denoted as the *feedback gain*. $G'_{d',\Delta}$ contributes for the gain-controlled noise source to model the quantizer with SNR_0 as defined in (5.10) so that

$$E\{\Delta^2(k)\} = \sigma_\Delta^2 = \frac{E\{d'^2(k)\}}{\mathrm{SNR}_0}. \tag{5.53}$$

The LP coefficients are assumed to be constant since $x(k)$ is stationary ($H_0(z)$ is constant). Therefore also all filtering gains are assumed to be constant.

5.2.3 Computation of the overall coding SNR ($\mathrm{SNR_{lpc}}$)

For the determination of the overall SNR according to (5.45), $E\{(\tilde{x}(k)-x(k))^2\}$ shall be computed as a function of the variance of the signal to be coded $E\{x^2(k)\}$. With respect to the signal generation model, $E\{x^2(k)\}$ is given as

$$E\{x^2(k)\} = E\{d_0^2(k)\}\cdot G_{d_0,x} = \sigma_{d_0}^2 \cdot G_{d_0,x} \tag{5.54}$$

Based on this, the variance of the LP residual signal $d(k)$ can be written as

$$E\{d^2(k)\} = E\{d_0^2(k)\}\cdot G_{d_0,d} = \frac{E\{x^2(k)\}}{G_{d_0,x}}\cdot G_{d_0,d}. \tag{5.55}$$

The signals $e_f(k)$ and $d(k)$ are assumed to be uncorrelated (refer to Section 5.2.5). Therefore, the corresponding variances are added to produce the variance of signal $d'(k)$,

$$E\{d'^2(k)\} = E\{d^2(k)\} + E\{e_f^2(k)\}. \tag{5.56}$$

The relation between the variance of the quantization error $\Delta(k)$ and its filtered version $e_f(k)$ is given as

$$E\{e_f^2(k)\} = E\{\Delta^2(k)\} \cdot G_{\Delta,e_f}, \tag{5.57}$$

and the variance of the quantization error inherent to the decoder output signal as

$$E\{(\tilde{x}(k) - x(k))^2\} = \sigma_\Delta^2 \cdot G_{\Delta,\tilde{x}-x}. \tag{5.58}$$

Substituting (5.53) in (5.57) in the first and the result in (5.56) in the second step yields

$$E\{d'^2(k)\} = E\{d^2(k)\} \cdot \frac{1}{1 - \frac{G_{\Delta,e_f}}{\mathrm{SNR}_0}}, \tag{5.59}$$

and, with (5.55) and (5.58), (5.59) can be written as

$$E\{(\tilde{x}(k) - x(k))^2\} = E\{x^2(k)\} \cdot \frac{G_{\Delta,\tilde{x}-x}}{\mathrm{SNR}_0} \cdot \frac{G_{d_0,d}}{G_{d_0,x}} \cdot \frac{1}{1 - \frac{G_{\Delta,e_f}}{\mathrm{SNR}_0}}. \tag{5.60}$$

From this equation the overall coding SNR as defined in (5.45) can finally be derived as

$$\mathrm{SNR}_{\mathrm{lpc}} = \frac{E\{x^2(k)\}}{E\{(\tilde{x}(k) - x(k))^2\}} = \frac{G_{d_0,x}}{G_{\Delta,\tilde{x}-x} \cdot G_{d_0,d}} \cdot (1 - \frac{G_{\Delta,e_f}}{\mathrm{SNR}_0}) \cdot \mathrm{SNR}_0. \tag{5.61}$$

5.2.4 Evaluation of the Noise Propagation Model

By defining different constraints for $H_A(z)$, $H_S(z)$ and $F(z)$, the noise propagation model can be configured for open- and closed-loop quantization. The open-loop case for $F(z) = 0$ has been investigated in [KV05]. As an outcome it is shown that open-loop quantization can benefit from correlation in the signal to be quantized by only partially decorrelating signal $x(k)$ in the LP analysis filter. The highest SNR is achieved by a "half-whitening" LP analysis filter which, however, is lower than the SNR achievable by closed-loop quantization.
For the generalized closed-loop quantization with the choice of $F(z) \neq 0$, the LP analysis filter is commonly configured to decorrelate the input signal as much as possible, hence $H_A(z) \approx H_0(z)$. A system stability constraint is derived as follows: By definition, the variance of the quantization noise in the decoder output must be positive. According to (5.60), however, this variance is only positive if

$$\frac{G_{\Delta,e_f}}{\mathrm{SNR}_0} < 1.0. \tag{5.62}$$

Therefore, the overall system is only stable if this constraint is fulfilled. If the stability constraint is not fulfilled, an unstable feedback loop evolves:

- The variance of signal $e_f(k)$ increases if the variance of the quantization error $\Delta(k)$ increases.
- The variance of the quantization error $\Delta(k)$ increases if the variance of signal $d'(k)$ increases since the quantizer produces a constant SNR_0.
- The variance of signal $d'(k)$ increases if the variance of signal $e_f(k)$ increases.

The resulting quantization error asymptotically has infinite variance[3].
If the system is stable the overall SNR for closed-loop LPC with respect to the definition of the filtering gains (5.48-5.52) is

$$\mathrm{SNR}_{\mathrm{lpc}} = \frac{G_{d_0,x}}{G_{\Delta,\tilde{x}-x} \cdot G_{d_0,d}} \cdot \left(1 - \frac{G_{\Delta,e_f}}{\mathrm{SNR}_0}\right) \cdot \mathrm{SNR}_0 \tag{5.63}$$

$$= \frac{\frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{1}{H_0(\Omega)}|^2 \mathrm{d}\Omega}{\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\Omega \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\Omega} \cdot \left(1 - \frac{\frac{1}{2\pi} \int_{-\pi}^{\pi} |F(\Omega)|^2 \mathrm{d}\Omega}{\mathrm{SNR}_0}\right) \cdot \mathrm{SNR}_0. \tag{5.64}$$

This equation shall in the following be discussed for different application constraints.

5.2.4.1 Methodology

For the evaluation of the derived noise propagation model, assumptions must be made about the correlation immanent to the signal to be quantized. In the new noise propagation model, the correlation of $x(k)$ can be controlled by defining sets of filter coefficients $\mathbf{a}_{\mathrm{ar}}$ to configure the AR filter with system function $H_0(z)$ (5.46) related to the AR process signal generation block. In the literature, most assumptions about signal correlation are based on AR processes of order $N_{\mathrm{ar}} = 1$ (AR(1) with $H_0(z) = 1 - \rho \cdot z^{-1}$) and the coefficient ρ close to 1.0, e.g. [T.C06]. Realistic signals, however, are much more complex than this. From narrowband speech coding it is well known that an LP predictor of order ten is a good choice to achieve a reasonable decorrelation of speech. Therefore, the correlation of speech should at least be modeled by an AR-process of order ten (AR(10)). Besides a realistic order of the AR process, realistic sets of AR filter coefficients $\mathbf{a}_{\mathrm{ar}}$ (to configure $H_0(z)$) are required to simulate typical characteristics of correlation in audio signals.
The way to achieve realistic results which was followed in this thesis is to analyze short-term stationary signal segments of real audio signals to produce example sets of AR filter coefficients as the basis for $H_0(z)$. Two such example sets of order 18 are listed in Appendix B.1 as $\mathbf{a}_{\mathrm{ar},1}$ and $\mathbf{a}_{\mathrm{ar},2}$. Given a set of AR filter coefficients and assuming that the input signal is perfectly decorrelated by the LP analysis filter ($H_A(z) = H_0(z)$) most system relevant parameters for an evaluation of the theoretical investigations can be easily computed. In that context the filtering gains

[3] This is obvious from (5.60) especially for $\frac{G_{\Delta,e_f}}{\mathrm{SNR}_0} = 1$.

Set	$G_{d_0,x}$ in dB	$G_{d_0,d}$ in dB	γ	G_{Δ,e_f} in dB	$G_{\Delta,\tilde{x}-x}$ in dB
$\mathbf{a}_{\mathrm{ar},1}$	19.62 dB	0 dB	1.0	15.46 dB	0 dB
			0.9	12.14 dB	2.45 dB

Table 5.1: Model parameters for example fixed set of AR filter coefficients $\mathbf{a}_{\mathrm{ar},1}$ and for different configurations of the error weighting filter $F(z) = A(z/\gamma)$ with $\gamma = 1.0$ and $\gamma = 0.9$.

in (5.48-5.52) are computed from the approximation of the corresponding magnitude spectra by means of a (long) Discrete Fourier Transform (DFT). Since the DFT spectrum is not continuous, the integral in (5.43) is approximated by the sum over the DFT coefficients. For one exemplary set of coefficients, $\mathbf{a}_{\mathrm{ar},1}$, all relevant model parameters are listed in Table 5.1 for $F(z) = A(z/\gamma)$ with $\gamma = 1.0$ (closed-loop quantization without noise shaping) and $\gamma = 0.9$ (closed-loop quantization with moderate noise shaping). $\mathbf{a}_{\mathrm{ar},1}$ was chosen as an example here as it well demonstrates the behavior of the new model and the corresponding different results compared to those related to the conventional theory of LPC and at the same time is very typical for segments of audio signals.

5.2.4.2 Closed-loop Quantization for High Bit Rates

At first, for the evaluation of the LPC model for closed-loop quantization, high bit rates are assumed and the error weighting filter is set to $F(z) = A(z/\gamma) = A(z)$ with $\gamma = 1.0$ (no noise shaping). Due to the high bit rate, the quantization error signal variance can be assumed to be very low, hence $\mathrm{SNR}_0 \gg G_{\Delta,e_f}$ for realistic signals. In this case, the overall coding SNR is

$$\mathrm{SNR}_{\mathrm{lpc,hr}} = G_{d_0,x} \cdot \mathrm{SNR}_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{1}{H_0(\Omega)}|^2 \mathrm{d}\Omega \cdot \mathrm{SNR}_0 \tag{5.65}$$

since $1 - \frac{G_{\Delta,e_f}}{\mathrm{SNR}_0} \approx 1$, $G_{d_0,d} = 1$, and $G_{\Delta,\tilde{x}-x} = 1$ in (5.63). Taking into account the zero-mean property of the AR filter $H_0(z)$ in the signal generation process (Section 5.1.5), (5.65) can be written as

$$\mathrm{SNR}_{\mathrm{lpc,hr}} = \underbrace{\frac{\frac{1}{2\pi} \int_{-\pi}^{\pi} |\frac{1}{H_0(\Omega)}|^2 \sigma_{d_0}^2 \mathrm{d}\Omega}{\exp\left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln(|\frac{1}{H_0(\Omega)}|^2 \sigma_{d_0}^2) \mathrm{d}\Omega \right)}}_{=G_p} \cdot \mathrm{SNR}_0. \tag{5.66}$$

The first part of the term on the right hand side is known as the *maximum prediction gain* G_p, expressed as the inverse of the spectral flatness measure (SFM) [MG76] of

signal $x(k)$,

$$\frac{1}{G_p} = \Xi_{\mathrm{SF}}(x(k)) = \frac{\exp\left(\frac{1}{2\pi}\int\limits_{-\pi}^{\pi} \ln(|\frac{1}{H_0(\Omega)}|^2\sigma_{d_0}^2)\mathrm{d}\Omega\right)}{\int\limits_{-\pi}^{\pi} |\frac{1}{H_0(\Omega)}|^2\sigma_{d_0}^2\mathrm{d}\Omega}. \tag{5.67}$$

The resulting logarithmic SNR in dB is

$$10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,hr}})_{\gamma=1.0} = 10 \cdot \log_{10}(G_p) + 10 \cdot \log_{10}(\mathrm{SNR}_0) \tag{5.68}$$

which is the result known from the literature for high bit rate assumptions, e.g., [MG76], [GR92], [VM06]. The correlation of the signal to be quantized can hence be transformed into a benefit with respect to the overall coding SNR. Both parts in (5.68) are independent from each other so that this result is a good motivation afterwards to develop a quantizer optimized for memoryless sources independently from its application in the context of LPC. The term (5.68) is consistent with rate distortion theory (2.50).
In the next step, the new model is configured for closed-loop quantization with noise shaping, that is, to spectrally shape the quantization noise in the decoder output to increase the perceived coding quality with respect to the properties of the human auditory system, e.g., [VM06]. For this purpose, the error weighting filter is chosen as $F(z) = A(z/\gamma)$ with $\gamma = 0.9$. However, it is well-known that the choice of $\gamma < 1.0$ also reduces the overall coding SNR. This is visible also in Table 5.1, since the filtering gain $G_{\Delta,\tilde{x}-x}$ is increased in comparison to the closed-loop LPC without noise shaping ($\gamma = 1.0$) due to the chosen value of γ. The overall logarithmic SNR is hence

$$\begin{aligned} 10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,hr}})_{\gamma=0.9} &= 10 \cdot \log_{10}(G_p) + 10 \cdot \log_{10}(\mathrm{SNR}_0) - \log_{10}(G_{\Delta,\tilde{x}-x}) \\ &< 10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,hr}})_{\gamma=1.0} \end{aligned} \tag{5.69}$$

5.2.4.3 Closed-loop Quantization for Low Bit Rates

For the theoretical evaluation of a low bit rate scenario, the quality of the quantizer is assumed to be

$$10 \log_{10}(\mathrm{SNR}_0) = 16 \text{ dB}, \tag{5.70}$$

According to the **conventional high rate theory**, given the example set of filter coefficients $\mathbf{a}_{\mathrm{ar},1}$ with the filtering gains as given in Table 5.1 and the assumption of the quality of the quantizer according to (5.70), the overall logarithmic SNR according to the high rate theory (5.68) would be

$$10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,hr}})_{\gamma=1.0} = 35.62 \text{ dB} \tag{5.71}$$

for $\gamma = 1.0$ and

$$10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,hr}})_{\gamma=0.9} = 33.17 \text{ dB}. \tag{5.72}$$

for $\gamma = 0.9$ (5.69).
For the analysis of closed-loop quantization at low bit rates based on the **new model**, at first, the choice of $\gamma = 1.0$ for the error weighting filter $F(z) = A(z/\gamma)$ is considered. Given the filtering gains in Table 5.1 for the example set of filter coefficients $\mathbf{a}_{\mathrm{ar},1}$ and the quality of the quantizer according to (5.70), the overall system is stable as the stability constraint (5.62) is fulfilled, $G_{\Delta,e_f}/\mathrm{SNR}_0 = 0.883 < 1.0$, and the overall logarithmic SNR according to the new model is

$$10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,lr}})_{\gamma=1.0} = 26.28 \text{ dB}. \tag{5.73}$$

This is significantly lower than the value as predicted according to the conventional theory on LPC for high bit rates (5.71).
If the new model is configured for closed-loop quantization with the error weighting filter computed as $F(z) = A(z/\gamma)$ with $\gamma = 0.9$, according to Table 5.1, besides the modified filtering gain $G_{\Delta,\tilde{x}-x}$, the feedback gain G_{Δ,e_f} is reduced. As a result, the overall logarithmic SNR according to the **new model** with the filtering gains as given in Table 5.1 for the example set of filter coefficients $\mathbf{a}_{\mathrm{ar},1}$ and the quality of the quantizer according to (5.70), is

$$10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,lr}})_{\gamma=0.9} = 30.86 \text{ dB} > 10 \cdot \log_{10}(\mathrm{SNR}_{\mathrm{lpc,lr}})_{\gamma=1.0}. \tag{5.74}$$

and hence higher than the SNR for $\gamma = 1.0$. In addition, the value is lower than the value according to the theory of LPC known from the literature (5.69). As a conclusion, according to our new model and in contrast to the results as given for high bit rates, the choice of $\gamma < 1.0$ is not only beneficial due to psychoacoustical reasons but also increases the overall coding SNR for low bit rates.

5.2.4.4 Encoder Stabilization by Noise Shaping

Considering a configuration with another quantizer with a quantization SNR of

$$10 \cdot \log_{10}(\mathrm{SNR}_0) = 13 \text{ dB} \tag{5.75}$$

compared to the value of 16 dB as before (5.70), the overall system would be unstable for $\gamma = 1.0$ and set $\mathbf{a}_{\mathrm{ar},1}$ since the stability constraint (5.62) would no longer be fulfilled. In this case the choice of $\gamma = 0.9$ would be the solution to stabilize the complete system since the feedback gain G_{Δ,e_f} is reduced, and (5.62) is fulfilled.

5.2.4.5 Measurements of Closed-loop Quantization Result

In order to verify the model and validate the new noise propagation model, the theoretical results were confirmed by measurements in the context of a real LP based coding scheme following the diagram in Figure 5.4 b). Compared to the results from the previous evaluations, a practical confirmation of the overall SNR (5.10) requires to produce a stationary signal with the same statistical properties as given by the definition of AR filter coefficients in the model. For this reason, an artificial signal was produced based on the set $\mathbf{a}_{\mathrm{ar},1}$ of filter coefficients for $H_0(z)$ and a Gaussian

noise excitation $d_0(k)$ with unit variance and zero mean as the output of an AR process following the principle of the signal generation block in Figure 5.7. The artificial signal was quantized and the overall SNR measured afterwards for both considered scenarios, $F(z) = A(z/\gamma)$ with $\gamma = 1.0$ and $\gamma = 0.9$. In order to realize the quantizer with a constant quantization SNR_0 (5.10), a logarithmic SQ according to Section 3.1.1.5 was employed. The compressor curve of a logarithmic SQ in practice has a logarithmic and a linear part. If signal amplitudes to be quantized fall into the linear part, SNR_0 would no longer be constant. In order to avoid that this falsifies the measurements, the quantizer was realized as a mathematical rule,

$$\tilde{d}(k) = Q(d''(k)) = \mathrm{sign}(d''(k)) \cdot \exp(C_{\log} \cdot \left\lfloor \frac{\log(|d''(k)|)}{C_{\log}} + 0.5 \right\rfloor). \tag{5.76}$$

Naturally, a finite bit rate for this type of quantizer can not be calculated since the quantizer resolution would become infinite for amplitudes approaching zero. Nevertheless, the measurement of the quantization SNR_0 for different values of the parameter $C_{\log}$ is sufficient for a validation of the noise propagation model since $\mathrm{SNR}_{\mathrm{lpc}}$ was given as a function of SNR_0 in all previous evaluations as well.
In parallel to the *measured* values for the overall $\mathrm{SNR}_{\mathrm{lpc}}$ (directly computed from the power of the input signal and the power of the quantization noise in the decoder output), the variances which are required to compute "real" values for filtering gains G_{Δ,e_f} and G_p were approximated by measurements of the signal powers of signal $x(k)$, $d(k)$, $\Delta(k)$ and $e_f(k)$, respectively. Based on these "real" values for filtering gains SNR values could be *computed* according to equation (5.63) to validate the theoretical results in comparison to the measured ones. The results are shown for the cases $F(z) = A(z/\gamma)$ with $\gamma = 1.0$ and $\gamma = 0.9$ in Figure 5.9. In addition to all that, also the SNR predicted by the high rate approximation in Section 5.2.4.2 (conventional high rate theory, (5.71)) and the values according to (5.74) and (5.73) (marked by the red bullets) are shown for $10 \cdot \log_{10}(\mathrm{SNR}_0) = 16$ dB for comparison. These values are slightly different than the measured values since the "real" filtering gains marginally deviate from those given in Table 5.1. The area of values $\mathrm{SNR}_0 < 15.5$ dB for which the model predicts that the overall encoder becomes unstable for $\gamma = 1.0$ is highlighted by the gray background color. In that area, the overall system is no longer linear.
The presented curves confirm that the new model is significantly more consistent with the measured results than the conventional theory of LPC and demonstrates that LPC does not benefit from the full prediction gain for lower bit rates.

5.2.5 Discussion of the Model

The new quantization noise production and propagation model is the basis for the generalization of the conventional high rate theory for LPC towards lower bit rates. In particular, it was shown that for lower bit rates, the overall logarithmic SNR is significantly lower than the value predicted by the conventional theory for high bit rates and that the closed-loop encoder (the combination of quantization,

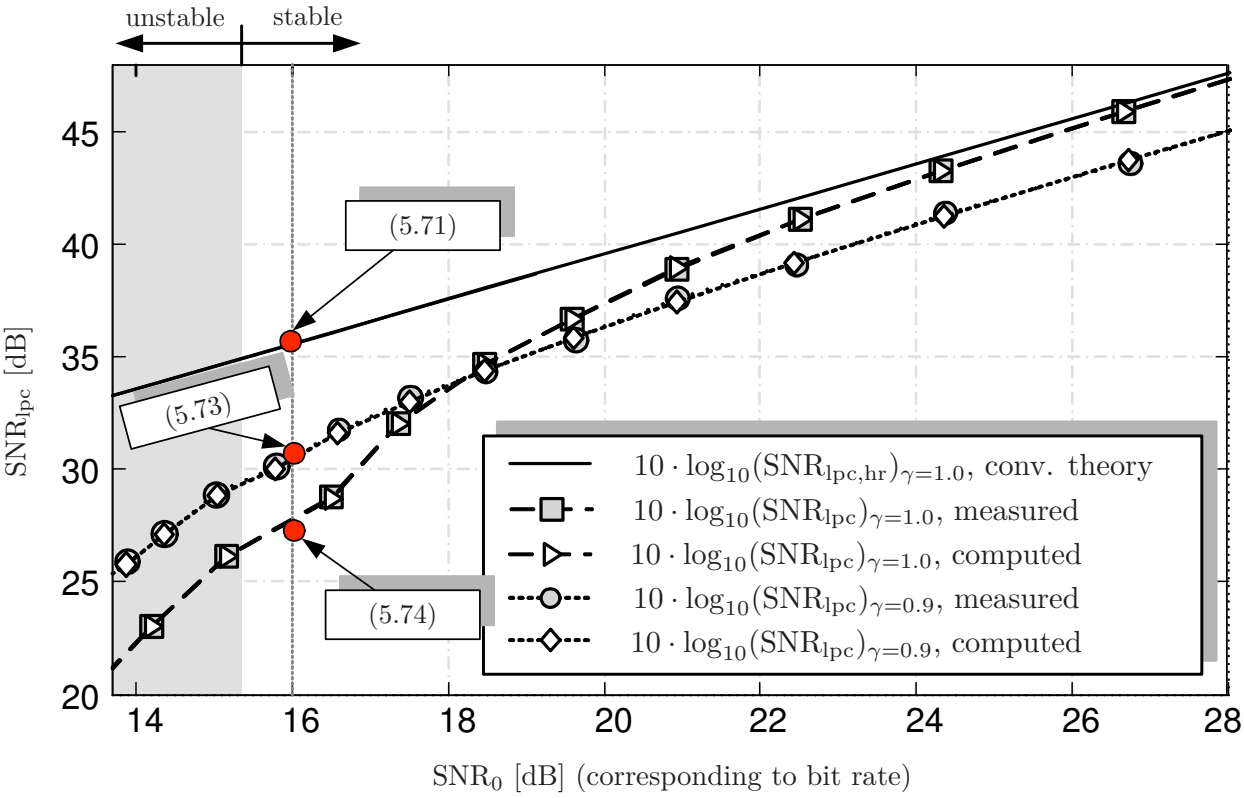


Figure 5.9: Logarithmic SNR in dB measured and predicted by the new model for $F(z) = A(z/\gamma)$ with $\gamma = 0.9$ and $\gamma = 1.0$. The SNR according to the conventional (high rate) theory of LPC is shown as a reference. Values for SNR_0 in which the overall encoder can become unstable for $F(z) = A(z/\gamma)$ with $\gamma = 1$ are highlighted by the gray background color. For higher bit rates, the achievable SNR is identical to what the conventional theory of LPC predicts. For lower bit rates, however, the quantization error feedback plays a more and more important role, and the computed overall SNR deviates from the SNR computed according to the conventional theory. The red bullets are the computed theoretical results according to (5.71), (5.74) and (5.73). Note that the system is no longer predictable in the (unstable operation) area $\mathrm{SNR}_0 \leq 15.5$ dB due to non-linear behavior.

noise feedback and linear prediction) can become unstable. Some open questions, however, still need to be discussed in the following.

5.2.5.1 Symptoms for Unstable Operation Conditions

It was described that the closed-loop LPC encoder can become unstable. This, however, is not clearly visible in Figure 5.9. So, what happens if the encoder is unstable? The answer is that nothing must happen but terrible things can happen. In quantization, in practice, the variance of the quantization error signal is a **mean value**. That means that quantization errors with smaller and with larger error amplitudes occur. The development of a noticeable artifact due to an instability in general has the duration of more than one sample interval if the stability constraint (5.62) is marginally unfulfilled. Therefore, in order to be noticeable as such, the instability requires that a sequence of large quantization errors occurs. Artifacts

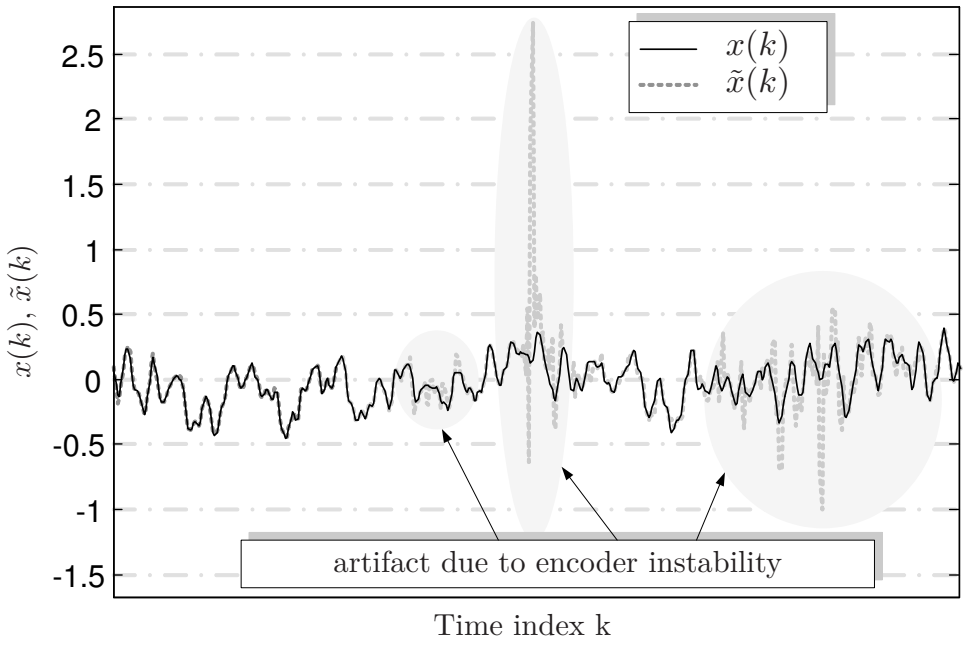


Figure 5.10: Example for artifacts as the result of an unstable operation condition.

develop more quickly if the feedback gain is high and are more likely to occur if SNR_0 is low.
An example for the typical symptoms of an instability is shown in Figure 5.10 for an example sequence of the artificial stationary signal from Section 5.2.4.5. The signals $x(k)$ and $\tilde{x}(k)$ are very much alike for a long time but suddenly a peak occurs. This peak is a very annoying acoustical artifact and is not tolerable. Unstable operating situations therefore should definitely be avoided. A rule of thumb to avoid instabilities is that the error feedback signal $e_f(k)$ should not have a higher power than signal $d(k)$ for a longer time. Unfortunately, it is hard to distinguish between wanted behavior (e.g., sudden decay of signal $d(k)$) and the development of instabilities just by observing the two signals in practice.

5.2.5.2 Relevance for Practical Coding Concepts

The evaluation results presented so far are based on reproducible, artificially generated quasi-stationary signals originating from an auto-regressive (AR) source model simulator. Practical audio signals are indeed significantly more complex than this. For example, quasi-stationarity is only present for short segments, and a lot of coding concepts exploit properties of types of signals to increase coding efficiency which are much more specific than the short-term correlation (e.g. the exploitation of long-term correlation in speech coding). So, are these results relevant for coding in practice then? The answer is yes: If already for very weak assumptions about the signals and under idealized conditions artifacts can occur, these artifacts definitely occur while quantizing real audio signals which are more complex either, together with other problems such as filter switching artifacts in the case of time variant LP synthesis filters etc.. The very simple signal model has helped to isolate the problem of instabilities in LPC with closed-loop quantization and to develop a theory. It

therefore indeed has a high significance for the development of the W-SCELP codec (Section 6).

5.2.5.3 Validity of the Model

A very important assumption made in the derivation of the new model was that two signals are uncorrelated, signal $d_0(k)$ as the excitation of the AR process and the quantization error $\Delta(k)$. Signal $d_0(k)$ by definition is uncorrelated as it is the output of an uncorrelated noise source. In order to model $\Delta(k)$ as a spectrally flat uncorrelated signal a bit rate which is sufficiently high is assumed, but the new model claims to be valid for lower bit rates also. Therefore, the assumption made here seems to be questionable at the first glance. Considering the results from Figure 5.9, however, it is pretty obvious that we have to distinguish between low bit rates for quantization in general and low bit rates in the context of coding of correlated signals: Starting from high values for SNR_0, the area in which the new theory starts to differ from the results according to the conventional theory begins at values of $\mathrm{SNR}_0 \approx 24$ dB. Taken into account the results from Section 3.1.1.5, a practical A-Law quantizer requires an effective bit rate of approximately 5-6 bits per sample to reach this (and approximately 4 bits to reach $\mathrm{SNR}_0 \approx 16$), and nobody would doubt that a quantizer still produces a quantization error which is independent of the input signal at these bit rates. As a conclusion, the bit rate at which the quantization error is no longer uncorrelated (the *low bit rate* in quantization in general) is much lower than the low bit rate at which the new model is valid and for which stability of the error feedback loop is given.

5.2.5.4 Conclusion of the Model

The definition of low bit rate areas in the context of the quantization of sources with memory has already been described in Section 2.3.4 in the context of the rate distortion theory. The new model confirms that the achievable performance gain due to an exploitation of correlation is lower for lower bit rates than the value predicted by the high rate asymptotic theory. In rate distortion theory, the loss toward lower bit rates was due to the impact of the reverse water filling. In this section, it was due to the feedback of quantization noise. The reverse waterfilling behavior is related to valleys of the spectrum of a signal (see Section 2.3.3), and it can be shown that mostly deep valleys in the spectrum $F(z = e^{j\Omega})$ of the error weighting filter are responsible for high feedback gains. Therefore and due to the following investigations, it will become more and more clear that, indeed, the reverse waterfilling is related to the feedback gain in LPC and that instabilities of the LPC encoder are a direct consequence of the disregard of the reverse waterfilling.
Two main conclusions can be derived from the evaluation of the new model that will be discussed in detail in the following:

- For lower bit rates, the quantizer and the surrounding signal processing should no longer be optimized independently in closed-loop LPC.

- One possible way to improve the system behavior of the LPC encoder is to employ the error weighting filter $F(z) = A(z/\gamma)$ with $\gamma < 1.0$ to be controlled adaptively in order to adapt to different signal types.

We can conjecture that, given a signal with a specific correlation, certainly there is an optimum γ to maximize the SNR. This optimal γ, unfortunately, cannot be calculated analytically. Nevertheless, an alternative method to determine the coefficients of the noise feedback filter $F(z)$ will be developed next.

5.2.6 A Novel Optimization Criterion for Closed-loop LPC

An important observation in the analysis of the new model in the previous section is that the signals $d(k)$ and $d'(k)$ in Figure 5.2 b) ($d'(k)$ computed from $d''(k)$ in Figure 5.6 for VQ) can significantly differ. With the assumption of a constant quantization SNR_0 produced by the quantizer, the quantization noise power in the decoder output signal is always proportional to the short-term power of signal $d'(k)$. Taking this fact into account, a new optimization criterion which targets the computation of the optimal LP coefficients for closed-loop quantization is

$$E\{d'^2(k)\} \rightarrow \min \tag{5.77}$$

which differs from the optimization criterion for linear prediction as proposed in the literature (5.3), $E\{d^2(k)\} \rightarrow \min$. Also, $H_A(z)$ and $F(z)$ are no longer assumed to necessarily employ the same or a similar set of filter coefficients. With a_i as the coefficients of the LP analysis filter system function $H_A(z)$ (5.2) and the coefficients b_i as the coefficients of the error weighting filter $F(z)$ (same order N_{lpc} in both cases), signal $d'(k)$ is computed as (see Figure 5.2 a))

$$d'(k) = x(k) - \underbrace{\sum_{i=1}^{N_{\mathrm{lpc}}} a_i \cdot x(k-i) - \sum_{i=1}^{N_{\mathrm{lpc}}} b_i \cdot \Delta(k-i)}_{\hat{\tilde{x}}(k)} . \tag{5.78}$$

A trivial solution for the problem from (5.77) is to set $b_i = 0$ for all $i = 1, \ldots, N_{\mathrm{lpc}}$ and to compute the a_i as in conventional linear prediction. This solution, however, is undesired since it is identical to **open-loop** quantization. A solution for **closed-loop** quantization can be derived by introducing the additional constraint to minimize the noise gain $G_{\Delta,\tilde{x}-x}$ in order to maximize the overall SNR. Unfortunately, this constraint can only be formulated in the frequency domain (5.51) since filter $H_S(z)$ is an IIR (infinite impulse response) filter.
An alternative solution can be realized as the following two-step-procedure:

- In the first step, the coefficients b_i of filter system function $F(z)$ and a_i of filter system function $A(z)$ are coupled by the relation $b_i = a_i$. Correspondingly, $F(z) = A(z)$ and the second constraint is perfectly fulfilled ($G_{\Delta,\tilde{x}-x} = 1$). Due to the modified optimization criterion, the resulting LP analysis is different to the conventional approach.

- In the second step, the coefficients a_i to be used in the LP analysis filter $H_A(z)$ are computed according to the conventional LP analysis.

The two steps are in detail realized and motivated as follows:

Step I: Modified LP Analysis

Signal $d'(k)$ is written as

$$d'(k) = x(k) - \sum_{i=1}^{N_{\text{lpc}}} b_i \cdot x(k-i) - \sum_{i=1}^{N_{\text{lpc}}} b_i \cdot \Delta(k-i). \tag{5.79}$$

Assuming that the signals $x(k)$ and $\Delta(k)$ are independent one from the other and that signal $\Delta(k)$ is uncorrelated (see Section 5.2.5.3), the new optimal filter coefficients are calculated from the following set of equations:

$$\begin{bmatrix} \varphi_{x,x}(1) \\ \dots \\ \varphi_{x,x}(N_{\text{lpc}}) \end{bmatrix} = (\mathbf{\Phi}_x + \mathbf{\Phi}_\Delta) \cdot \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ b_{N_{\text{lpc}}} \end{bmatrix} \tag{5.80}$$

with

$$\mathbf{\Phi}_x = \begin{bmatrix} \varphi_{x,x}(0) & \dots & \varphi_{x,x}(1-N_{\text{lpc}}) \\ \varphi_{x,x}(1) & \dots & \varphi_{x,x}(2-N_{\text{lpc}}) \\ \dots & \dots & \dots \\ \varphi_{x,x}(N_{\text{lpc}}-1) & \dots & \varphi_{x,x}(0) \end{bmatrix} \tag{5.81}$$

and

$$\mathbf{\Phi}_\Delta = \begin{bmatrix} \varphi_{\Delta,\Delta}(0) & \dots & 0 \\ 0 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \varphi_{\Delta,\Delta}(0) \end{bmatrix} \tag{5.82}$$

The first part of this equation involving $\mathbf{\Phi}_x$ is identical to the conventional approach in linear prediction, and the second part involving $\mathbf{\Phi}_\Delta$ is related to the feedback of the quantization error due to the error weighting filter $F(z)$. In matrix $\mathbf{\Phi}_\Delta$, the term $\varphi_{\Delta,\Delta}(0)$ depends on the performance of the quantizer, SNR_0, and the power of the signal $d'(k)$ which again depends on the coefficients b_i. (5.80) would hence no longer be a linear set of equations and a solution not straight forward.
A reasonable approximation of the quantization noise power, however, is to set it to a constant value $\varphi_{\Delta,\Delta}(0) = C_\Delta \cdot \varphi_{x,x}(0)$ for the computation of the coefficients b_i. The resulting approach is well-known as *white-noise-correction* (WNC) which, however, was introduced to avoid ill-conditioned autocorrelation matrices in the LP analysis, e.g., [KP95]. The corresponding new error weighting filter is denoted as

$$F_{\text{new}}(z) = \sum_{i=1}^{N_{\text{lpc}}} b_i \cdot z^{-1}. \tag{5.83}$$

Note that the order of $F_{\text{new}}(z)$ is not required to be equal to N_{lpc}.

Step II: Minimization of $d(k)$

In order to furthermore optimize the overall system performance, another conclusion is derived from the new noise production and propagation model: From (5.57) it is obvious that only $F(z)$ but not $H_A(z)$ has an impact on the variance of signal $e_f(k)$ and hence the term $(1 - G_{\Delta,e_f}/\text{SNR}_0)^{-1}$ in (5.60). Therefore, if the error weighting filter $F(z)$ is fixed, the maximum overall quantization performance is achieved by minimizing filtering gain $G_{d_0,d}$ (refer to (5.55) and (5.59).). In order to minimize $G_{d_0,d}$ the coefficients a_i should be computed according to the conventional LP analysis (to approximate $H_0(z)$ by $H_A(z)$ as good as possible) whereas the proposed white-noise-correction should only be considered for the computation of the coefficients b_i of the error weighting filter $F_{\text{new}}(z)$.

In a practical application, the two-step-procedure can be efficiently realized, e.g., by computing the two sets of LP coefficients in two subsequent executions of the Levinson Durbin algorithm. It is applicable for combined LP and SQ as well as for CELP coding since, with respect to (5.27) and Section 5.1.7.3, the error weighting filter $W(z)$ and the error weighting filter $F(z)$ are equivalent. Constant C_Δ still is an unknown parameter which should be adapted to the quantizers bit rate. In the context of the W-SCELP and the SCELP codec (Section 6), the best choice for C_Δ was determined based on the evaluation of simulations.
In order to better understand the impact of the proposed two-step procedure, it will be analyzed in the frequency domain for the exemplary spectrum related to a set of AR filter coefficients in the following. It will be shown that the proposed method involving the new error weighting filter $F_{\text{new}}(z)$ is the time domain approximation of the reverse waterfilling procedure according to the rate distortion theory (Chapter 2) and therefore produces a higher coding SNR than the method involving the conventional error weighting filter $F_{\text{conv}}(z)$ from (5.14).

5.2.6.1 Reverse Waterfilling according to the Rate Distortion Theory

The reverse waterfilling principle was introduced in the context of the optimal quantization of correlated signals according to the rate distortion theory in Section 2.3. In this section, the knowledge about this principle in general will be refreshed, and its realization in practice will be explained based on a qualitative plot of the magnitude spectrum of an exemplary signal. In the next section, it will then be described why the modified LP analysis from the previous section is the realization of reverse waterfilling in closed-loop LPC.
For the explanation of the reverse waterfilling principle based on an exemplary magnitude spectrum, it is more useful to consider only spectral envelopes of signals rather than the exact fine structure of a high resolution magnitude spectrum. In Figure 5.11 an exemplary log spectral envelope $\text{SE}_X(\Omega)$ is shown as the blue line with the triangle markers. Again, with respect to the motivation given in Section 5.2.5.2, the shown spectral envelope is related to a stationary signal that originates

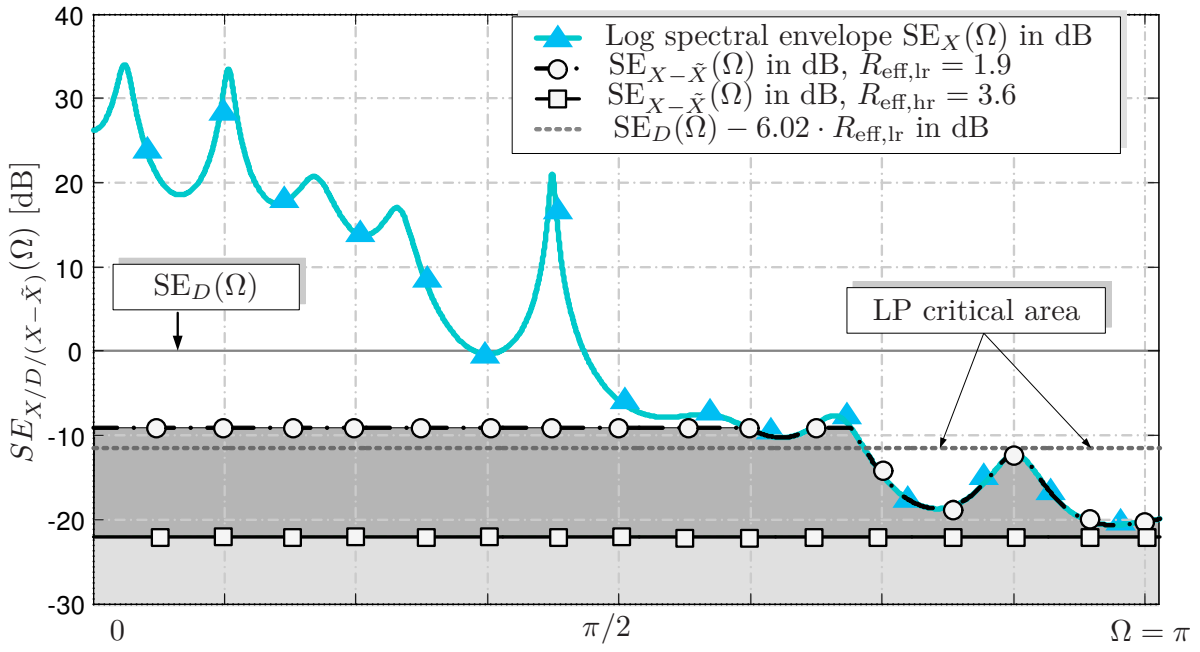


Figure 5.11: Reverse waterfilling principle according to the rate distortion theory illustrated for two example bit rates, $R_{\mathrm{eff},1} = 3.6$ bits per sample and $R_{\mathrm{eff},2} = 1.9$ bits per sample. All curves are constructed from the spectral envelope $\mathrm{SE}_X(\Omega)$ according to (5.84) for an exemplary set of AR filter coefficients in analogy to Section 5.2.4.1.

from an AR process. The spectral envelope is derived from the corresponding exemplary set of AR filter coefficients for the AR filter with system function $\frac{1}{H_0(z)}$ based on the substitution $z = e^{j\Omega}$, e.g. [VM06], as

$$\mathrm{SE}_X(\Omega) = 10 \cdot \log_{10}\left(\left|\frac{1}{H_0(\Omega)}\right|^2\right). \tag{5.84}$$

The log spectral envelope can be computed by means of a *Discrete Fourier Transform* (DFT) and is a normalized approximation of the logarithmic *Power Spectral Density* (PSD) which was the basis for the explanation of the reverse waterfilling in Section 2. It is normalized because the AR filter has zero-mean property (Section 5.1.5) which is not necessarily the case for the PSD. Since, according to the results from rate distortion theory, only relative level differences of variances are relevant, the normalization has no impact on the investigated issues. The zero-mean property has two consequences:

- $\mathrm{SE}_X(\Omega)$ is located around the 0-dB-line in Figure 5.11.
- A decorrelation of the signal, e.g., by means of LP analysis filtering, produces a signal with the normalized log spectral envelope

$$\mathrm{SE}_D(\Omega) = 0 \text{ dB}. \tag{5.85}$$

The shown example has low pass characteristic which is typical for audio signals. For the following explanations, two scenarios are considered to demonstrate the reverse waterfilling, one based on the assumption of a high effective bit rate, $R_{\mathrm{eff,hr}} =$

3.6 bits per sample, and one for a low effective bit rate of $R_{\text{eff,lr}} = 1.9$ bits per sample.

In the high bit rate example, the quantization is controlled according to the reverse waterfilling such that the quantization error variance is constant over frequency (2.50). Very generally, the quantizer is modeled by the 6-dB-per-bit rule (2.18) in Figure 5.11, so that the level of the quantization error in the decoder output, $\text{SE}_{X-\tilde{X}}(\Omega)$, is $10 \cdot \log_{10}(\text{SNR}_0) = 6.02 \cdot R_{\text{eff,hr}}$ dB lower than $\text{SE}_D(\Omega)$, that is, at $-6.02 \cdot R_{\text{eff,hr}} = -21.6$ dB on the y-axis. $\text{SE}_{X-\tilde{X}}(\Omega)$ is shown as the solid line with the square markers for the high bit rate in Figure 5.11.

In the low bit rate example, based on the 6-dB-per-bit rule (2.18) quantization model and the location of the spectral envelope $\text{SE}_D(\Omega)$, the spectral envelope of the quantization noise $\text{SE}_{X-\tilde{X}}(\Omega)$ could be expected to be located $10 \cdot \log_{10}(\text{SNR}_0) = 6.02 \cdot R_{\text{eff,lr}}$ dB lower than $\text{SE}_D(\Omega)$, that is, at $-6.02 \cdot R_{\text{eff,lr}} = -11.5$ dB on the y-axis for all Ω in analogy to the case for high bit rates. This curve is shown as the dotted line in the figure. However, the reverse waterfilling prescribes that the quantization noise level can never be higher than the level of the signal to be quantized (in rate distortion, this case is equivalent to 0 dB SNR). Considering this, the spectral envelope of the quantization noise is reduced in the areas denoted as the *LP critical areas* to follow the spectral envelope $\text{SE}_X(\Omega)$ (denoted as the "reduce" operation), and in all other areas, the quantization noise level is slightly raised (denoted as the "raise" operation). The resulting "correct" spectral envelope (according to RDT) is shown as the dash-dotted curve with the circle markers. The impact of the described procedure for lower bit rates is a reduced SNR compared to the SNR related to the dotted line. This behavior is also well documented in Section 2.3.4 by the SNR plots in Figure 2.4.

5.2.6.2 Reverse Waterfilling in Closed-loop Quantization

In closed-loop quantization the spectral envelope of the processed quantization noise in the decoder output is influenced by the weighting filter $F(z)$ according to the magnitude spectrum with respect to (5.13).

For the high bit rate example from the previous section, to configure the error weighting filter such that $A(z) = F(z)$ is well possible as feedback problems do not occur due to the fulfillment of the stability constraint (5.62). The spectral envelope of the quantization noise is constant for all Ω and located at the same position as in Figure 5.11 if the quantizer performance is assumed to follow the 6-dB-per-bit rule again.Since this case is trivial, it is not shown in the figure.

For the lower bit rate example, setting $A(z) = F(z)$ would lead to coding artifacts since the overall system is unstable due to the low quantizer SNR and the unfulfilled constraint (5.62). Instead, $F_{\text{new}}(z) \neq A(z)$ is computed according to the new optimization criterion from Section 5.2.6 and equation (5.77). As a conclusion, the log spectral envelope of the processed quantization noise in the decoder output

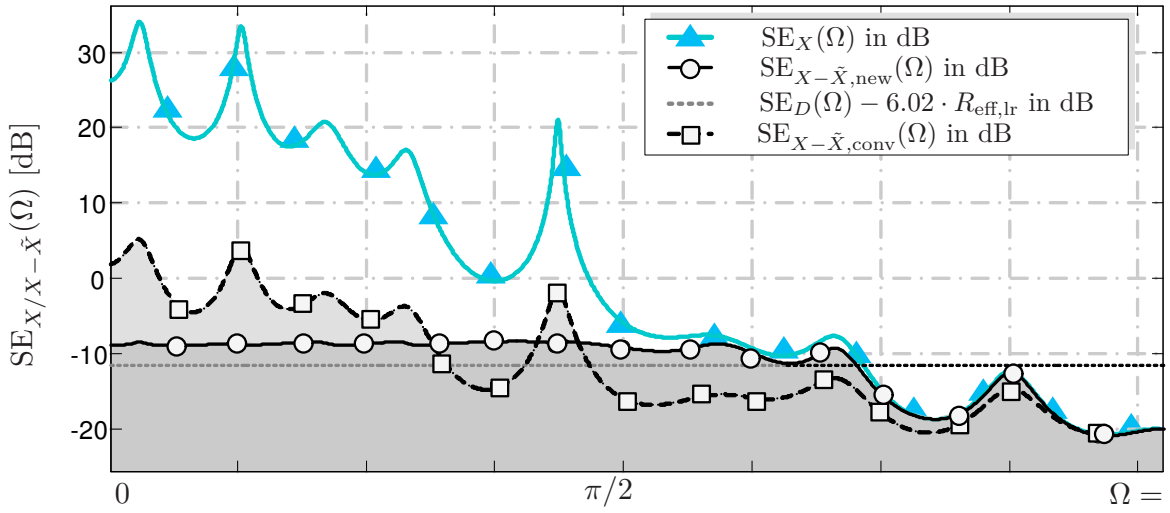


Figure 5.12: Reverse waterfilling in closed-loop quantization based on the filter $F_{\text{new}}(z)$ for the new approach (circle markers) and the conventional approach based on filter $F_{\text{conv}}(z)$ according to (5.14).

is

$$\text{SE}_{X-\tilde{X},\text{new}}(\Omega) = 10 \cdot \log_{10}\left(\left|\frac{1 - F_{\text{new}}(\Omega)}{1 - A(\Omega)}\right|^2\right), \tag{5.86}$$

shown in Figure 5.12 as the solid line marked by gray circles. Obviously, the spectral envelope very well approximates the optimal curve for the low bit rate example (dash-dotted curve with circle markers in Figure 5.11). The curves are not identical, though, since the frequency response is approximated by a time domain filter. It can be shown that the "raise" operation defined in the previous section is related to the white-noise-correction and hence the first step in the two-step procedure to compute $F_{\text{conv}}(z)$, whereas the "reduce" operation is related to the difference between $F_{\text{conv}}(z)$ and $A(z)$ and hence the second step.
In addition to the spectral envelope related to the choice $F_{\text{new}}(z)$, the log spectral envelope based on the approach proposed in the literature with a choice of the error weighting filter such that $F_{\text{conv}}(z) = A(z/\gamma)$ according to (5.14) is shown in Figure 5.12 as the dashed line with the square markers,

$$\text{SE}_{X-\tilde{X},\text{conv}}(\Omega) = 10 \cdot \log_{10}\left(\left|\frac{1 - F_{\text{conv}}(\Omega)}{1 - A(\Omega)}\right|^2\right), \tag{5.87}$$

In order to have a comparable impact as with $F_{\text{new}}(z)$, γ was chosen such that the spectral envelope of the processed quantization noise is below the spectral envelope of the input signal at all frequencies and in particular in the *LP critical areas* from Figure 5.11 ($\gamma = 0.55$).
As a conclusion of this diagram, the new approach approximates the spectral envelope which is optimal according to RDT much better than the conventional approach. The quantization noise level is significantly lower in the low frequency

areas, and the overall SNR is approximately 6 dB higher for the new approach than for the conventional approach for the example from the figure.
So far, in the discussion of LPC, effects related to human perception in audio coding were not considered since all investigations were solely based on the computations of SNRs. In order to take human perception into account in a practical coding scheme, the two concepts for the error weighting filter $F(z)$ can be combined. This and other aspects related to human perception will be subject of Section 6.
For the examples in Figure 5.12, the constants C_Δ and γ were tuned manually. In the application for coding, in general fixed values are used which are not optimal but offer a reasonable quality. In order to improve the concept, especially γ should be controlled adaptively for different segments of the input signal.

5.2.7 Adaptation of the Scalar Noise Model for CELP Coding

In the new noise propagation model for LPC the quantizer is modeled by a **scalar** additive noise source. The motivation to generalize the results from the scalar model also to CELP coding was explained in Section 5.1.7.3 where it was shown that closed-loop LPC with SQ and VQ are equivalent if $W(z)$ is configured according to (5.27). Also, it is clear that for the same effective bit rate per sample VQ has higher performance than an SQ due to the VQ advantages (Section 3.2.1). But besides these facts, are there any other differences between SQ and VQ in the context of the new model for closed-loop linear predictive coding, or, in other words, does the scalar model really also apply for CELP coding?
In order to find the answer to this question, the measurements from Section 5.2.4.5 were repeated involving a CELP encoder (based on the APVQ LSVQ candidate from Section 4.4.3 involving the testing of all codevectors in the codebook (full search) for quantization) to replace the logarithmic SQ and for $W(z) = 1$ (hence $F(z) = A(z)$ in the equivalent scalar model.). The measured results showed that, on the one side, the CELP encoder behaves as predicted by the derived new model: For lower bit rates, the quantization error signal is fed back due to the error weighting filter, the (equivalent) signal $d'(k)$ deviates from signal $d(k)$, the overall system can be unstable, and the error weighting filter $F_{\text{new}}(z)$ is also a good choice for VQ. On the other side, however, a CELP coder seems to be less sensitive against feedback problems than an SQ based approach.
The reason for this well-tempered behavior is that a CELP encoder benefits from the fact that more than one quantization reconstruction level is determined at once in the closed-loop codebook search. Due to the joint optimization of sequential samples, the search for the best codevector implies the introduction of a certain dependence between the samples of each quantized vector which can not be achieved by SQ. The introduction of dependence between samples has an impact similar to that of a filter. In order to **model** this filtering characteristic, the *implicit error weighting filter* $W'(z)$ is introduced in the CELP encoder model in Figure 5.13 and is responsible for a spectral shaping of the quantization noise. Consequently, $W'(z)$

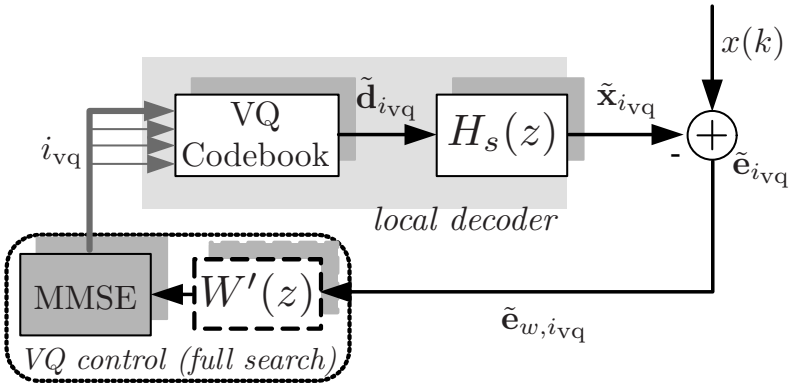


Figure 5.13: Basis for the extension of the scalar noise model for CELP coding: The encoder involving a full codevector search has the impact of an *implicit error weighting filter* $W'(z)$ which is due to the introduction of a specific dependence between subsequent samples in the quantized LP residual signal vectors caused by the full codevector search.

must also be considered in the equivalent scalar noise production model in order to be valid for CELP coding and helps to better match the reverse waterfilling to reduce the variance of signal $e_f(k)$. However, the impact of the implicit error weighting filter can only be validated indirectly:

- The measurements of signal powers during the simulation of CELP encoding with a choice of $W(z) = 1$ (and hence $F(z) = A(z)$ in the equivalent scalar model) show that the "real" filtering gain G_{Δ,e_f} differs significantly from the theoretical values derived from the filter coefficients of filter $F(z)$ (according to (5.51)). This behavior can only be explained by a quantization error $\Delta(k)$ which is not spectrally flat.

- The quantization noise in the decoder output is expected to be spectrally flat ($\mathrm{SE}_{X-\tilde{X},\mathrm{conv}}(\Omega) = \mathrm{const}$) for $W(z) = 1$. However, the measurements showed that this is not the case but, instead, the spectral shape was very similar to the spectral shape according to the reverse waterfilling procedure. This behavior can only be explained by the implicit error weighting filter $W'(z)$.

The impact of the implicit error weighting filter, however, is limited for VQ vector dimensions used in practice, similar to the effect of truncated impulse responses in FIR filters for the approximation of desired frequency responses. Therefore it is not a good replacement for the near-optimal weighting filter as proposed in Section 5.2.6 in practical applications and for realistic vector dimensions. In addition to that, the computational complexity for a full codevector search is prohibitive for very low dimensions already so that most optimized CELP quantization procedures can afford to test only a fraction of the overall number of vectors in the VQ codebook. In these cases the impact of the implicit error weighting is almost imperceptible.

5.2.8 Encoder Stabilization in Speech Coding

LPC has received much attention in the speech coding community during the last decades. The issue related to the feedback of the quantization error, however, has

not been addressed by many researchers. One reason is that speech signals do not have such extreme characteristics as audio signals (e.g., segments with very high prediction and feedback gains for artificial and low pass filtered audio signals) and that codec parameters such as the length of the LP analysis segments are configured in a different way in low delay audio coding (e.g., to fulfill the low delay constraints). Furthermore, a lot of common techniques are employed in state-of-the-art speech codecs which prevent the evolution of artifacts caused by unstable feedback loops. These techniques, however, were introduced for other reasons:

- *White-noise-correction* has been introduced to avoid ill-conditioned matrices [KP95]. It was shown that this technique is the most important approach to combat feedback problems and to maximize the SNR.
- *Spectral bandwidth widening* in the LP spectrum has been introduced to better match formant frequencies [KP95]. This technique also reduces the feedback gain in LPC and hence feedback problems.
- The weighting filter $F(z) = A(z/\gamma)$ was introduced to achieve a better perceptual speech quality [SAH79]. It was shown that a value of $\gamma < 1.0$ leads to performance benefits also with respect to the SNR as a (suboptimal) alternative to the white-noise-correction.
- Very large vector dimensions (e.g., $L_v = 40$ based on the interleaved concatenation of vectors as described in Section C.3 of the *supplement document* [Krü09] in the AMR speech codec [ETS00]) for the quantization of the LP residual signal [SA89] are employed. It was explained that VQ with large vector dimensions also reduces the feedback gain due to the implicit error weighting filter.
- Longer LP analysis segments lead to moderate prediction and, in particular, feedback gains in conventional speech coding compared to low delay audio coding.

Tests based on the generic state-of-the-art JMEDIA ACELP speech codec developed in [KV02] showed that also speech coders can become unstable due to error feedback effects if the mentioned common techniques are not employed.

5.3 Discussion

In this section, linear predictive coding was investigated as an approach to exploit linear correlation within audio signals to increase the coding performance. In the first part, the principle of linear prediction was briefly reviewed. It was motivated that LPC is useful for speech signals as well as for audio signals and a good choice for low delay source coding. Combined closed-loop linear prediction and SQ was explained, and it was shown that this concept can be generalized as analysis-by-synthesis VQ which is denoted also as Code-Excited Linear Prediction (CELP).

Targeting the reduction of the complexity related to the CELP encoder, a modified encoder structure was developed which will be the basis for a significant reduction of complexity of the CELP encoder to be investigated in Section 6.1.3.
In the second part, novel aspects related to source coding at lower bit rates were investigated which are of high relevance for low delay audio coding as well as for a deeper fundamental understanding of LPC in general. A new model which, in contrast to the conventional theory of LPC given in the literature, is valid for high as well as for low bit rates showed that the interaction between the quantizer and the feedback of the quantization error can lead to encoder instabilities and overall performance losses. In order to combat these undesired effects, a novel optimization criterion was introduced to enable the computation of the error weighting (noise feedback) filter $F_{\mathrm{new}}(z)$ for closed-loop linear predictive quantization. It was shown that this novel approach much better considers the reverse waterfilling known from rate distortion theory than the conventional error weighting filter known from the literature.

6

The SCELP and the W-SCELP Low Delay Audio Codecs

In Section 4, the theory and concepts for practical realizations of LSVQ were developed. In Section 5, the combination of linear prediction and LSVQ for linear predictive coding (LPC) was investigated. In this chapter, the results from the two chapters shall be combined in order to develop the Spherical Code-Excited Linear Prediction (SCELP) low delay audio codec. Special attention is at first drawn to practical aspects, in particular to achieve a low computational complexity of the SCELP encoder. It will be shown that a huge reduction of computational complexity is possible by exploiting properties of the LSVQ codevectors with tolerable degradation of the coding performance.
In order to realize audio coding with high perceived quality, also design aspects related to human perception will be briefly discussed in the next step. As a result, the SCELP codec is enhanced by means of Warped Linear Prediction (WLP), denoted as the W-SCELP codec [KV07b].
In the last part of this chapter, results for the W-SCELP audio codec based on objective audio quality measurements will be presented. It will be shown that the SCELP low delay audio codec outperforms the ITU-T G.722 audio codec [ITU88b].

6.1 The SCELP Low Delay Audio Codec

The SCELP low delay codec was developed as a candidate for a digital solution for wireless audio-links in hearing aids as described in Section 1.1.2. Accordingly, low algorithmic delay, computational complexity and high perceived audio quality as well as transmission robustness were important constraints that needed to be fulfilled. It was explained in previous sections that LP combined with LSVQ is a promising candidate for audio coding with low algorithmic delay. In this section,

the computational complexity and the achievable audio quality will be investigated. The aspect of transmission robustness is discussed in detail in [KSV06] and shall not be examined here.
The computational complexity in CELP coding is mainly an issue of the encoder and, in particular, of the weighted codevector search procedure according to (5.24). The methods for complexity reduction discussed in the following are therefore based on simplifications of this functional part.
But why would the complexity be an issue here once more if highly efficient approaches for nearest neighbor quantization were proposed earlier for the LSVQ concepts? The answer is that it definitely is a big practical issue since the proposed nearest neighbor quantization approaches can not be reused due to the weighting matrix $\mathbf{H}_W$ in (5.24). All investigations about the theoretically achievable coding performance related to CELP coding in the previous chapters were based on the assumption that all codevectors in the LSVQ codebook are tested in analysis-by-synthesis manner as described in Section 5.1.7. This approach, denoted as the *full search* strategy in the following, is prohibitive already for low dimensions and bit rates. Therefore, the overall concept of combined linear prediction and LSVQ presented in this thesis would be of no practical use if a realization of the codevector search procedure with low complexity would be impossible.
Other researchers have identified this issue before: In the literature, efficient search strategies for combined linear prediction and VQ were in particular published in the context of the algebraic codebooks for speech coding, denoted as Algebraic Code-Excited Linear Prediction (ACELP), e.g., [LAS+91], [SLAM94]. The published strategies exploit the sparseness of the codebooks to achieve a high performance and a reasonable computational complexity at the same time. It was shown in Section 4.4.1, however, that this type of VQ is only useful to achieve moderate quality for very low bit rates, e.g., $10 \cdot \log_{10}(\mathrm{SNR}_0) = \mathrm{SNR}_{\mathrm{svq}}^{(II)}|_{\mathrm{dB}} < 7$ dB for bit rates of $R_{\mathrm{eff}} < 1.5$ bits per sample for the ALBVQ as shown in Figure 4.23[1]. A reuse of the proposed techniques for low delay audio coding with higher quality constraints is hence not possible. Besides this, theoretical or more general publications on efficient weighted vector search procedures, unfortunately, are barely available.
The other approaches from Chapter 4, the GLCVQ and the APVQ, were developed for the quantization of memoryless sources at higher bit rates ($R_{\mathrm{eff}} > 2$ bits per sample to achieve $10 \cdot \log_{10}(\mathrm{SNR}_0) = \mathrm{SNR}_{\mathrm{svq}}^{(II)}|_{\mathrm{dB}} > 10$ dB) to enable higher quality VQ. In order to employ these concepts also for CELP coding, new efficient search strategies for the weighted vector search had to be developed. Among the two concepts, the GLCVQ has the higher performance as documented in Section 4.4, and a nearest neighbor quantization procedure can be realized with low computational complexity (refer to Section D.2 of the *supplement document* [Krü09]). Unfortunately, all attempts to develop a low complexity version of the weighted codevector search in the context of CELP coding failed since the resulting loss of quality com-

[1] Indeed, the proposed efficient codevector search methods assume that only very few pulses are non-zero. Correspondingly, the number of pulses in the standardized speech codecs is very low so that in ACELP coding, the quality of the base quantizer is even lower.

Standard configuration of the SCELP codec					
Effective bit rate in bits $R_{\text{eff,lsvq}}$	Vector Dim. L_v	Sample-rate in Hz f_s	Number of spherical codevectors N_{svq}	Number of A-Law quantizer levels N_g	Overall number of LSVQ codevectors $N_{\text{lsvq}} = N_g \cdot N_{\text{svq}}$
2.0	11	22050	105328	27	2843856

Table 6.1: Standard configuration of the SCELP low delay audio codec to illustrate the possible reduction of complexity and corresponding quantization performance loss.

pared to a full codevector search was unacceptable [Sch06]. Therefore, an efficient realization of the codevector search for combined CELP coding and LSVQ was developed for the APVQ from Section 4.4.3 as well. The APVQ was shown to have only a marginally lower performance than the GLCVQ for memoryless sources. It is the better choice for CELP coding, however, since the underlying Apple Peeling spherical code enables to realize a reduced complexity codevector search to save huge amounts of complexity while at the same time decreasing the quality only marginally compared to the full search approach.
The highly efficient combination of CELP and APVQ forms the SCELP codec and is explained in detail in [KV06b]. In this section, only a brief summary of the results is presented which is derived from more detailed investigations about efficient codevector search procedures in CELP coding based on principal considerations. Detailed measurements for artificial quasi-stationary signals and visualization of quantization cell shapes for a simple two-dimensional VQ are presented in Section E of the *supplement document* [Krü09].

6.1.1 The SCELP Standard Configuration

The SCELP codec is applicable for a wide variety of applications. In order to illustrate the efficiency of the proposed techniques, the reduction of the computational complexity and the corresponding loss of quantization performance due to the reduced codevector search effort shall be demonstrated by a concrete example:
In its standard configuration (all relevant parameters of the *SCELP standard configuration* are summarized in Table 6.1), the SCELP low delay audio codec is operated with an effective bit rate of $R_{\text{eff,lsvq}} = 2.0$ bits per sample, a vector dimension of $L_v = 11$, and a sample rate of $f_s = 22050$ Hz. For the given effective bit rate and under consideration of the optimal bit allocation for LSVQ from Section 4.2.4.4, the number of codevectors located on the sphere surface is $N_{\text{svq}} = 105328$, and the number of quantization reconstruction levels for the gain factor $N_g = 27$. The computation of the LP coefficients and the filter coefficients of the error weighting filter $F(z)$ was realized according to the approach proposed in Section 5.2.6 to contribute for the novel optimization criterion for linear predictive quantization with

error feedback from (5.77). All other parameters of the SCELP encoder such as, e.g., the LP order do not have a significant impact on the computational complexity of the CELP codevector search procedure and are therefore not listed in Table 6.1. The SCELP standard configuration is also the basis for the investigations in Section E.1 of the *supplement document* [Krü09]:

6.1.2 Maximum Theoretical Complexity and the Definition of a Quality Loss Measure

For a comparison of different codevector search strategies in CELP coding it is very important to know the maximum achievable performance given unlimited computational power. The maximum SNR is achieved by following the full search approach for the determination of the optimal codevector, denoted as S_{FS}. The involved computational complexity, indeed, is very high and prohibitive for practical applications:
Given the SCELP standard configuration listed in Table 6.1, according to the modified CELP approach from Section 5.1.7.1, metric

$$\mathcal{M}_{i_{\mathrm{lsvq}}} = (\mathbf{x}'_0 - \mathbf{H}_W \cdot \tilde{\mathbf{d}}_{i_{\mathrm{lsvq}}})^T \cdot (\mathbf{x}'_0 - \mathbf{H}_W \cdot \tilde{\mathbf{d}}_{i_{\mathrm{lsvq}}}). \tag{6.1}$$

must be computed for each of the $N_{\mathrm{lsvq}} = N_{\mathrm{svq}} \cdot N_g = 2843856$ LSVQ codevectors. The optimal codevector index $i_{Q,\mathrm{lsvq}}$ is determined by metric comparison in analogy to (5.24) as

$$i_{Q,\mathrm{lsvq}} = \underset{0 \le i_{\mathrm{lsvq}} < N_{\mathrm{lsvq}}}{\arg\min} \; \mathcal{M}_{i_{\mathrm{lsvq}}}. \tag{6.2}$$

The convolution of one LP residual vector candidate $\mathbf{d}_{i_{\mathrm{lsvq}}}$ with the truncated impulse response in $\mathbf{H}_W$ is assumed to be realized in $(L_v{+}1){\cdot}L_v/2$ multiply-accumulate instructions. Since the VQ is composed of spherical codevectors and gain factors, in order to construct $\mathbf{d}_{i_{\mathrm{lsvq}}}$, each (spherical) vector coordinate must be multiplied with a quantized gain factor which adds L_v multiply operations per candidate vector, and the evaluation of the quantization error is realized in L_v subtract instructions. The computation of $\mathcal{M}_{i_{\mathrm{lsvq}}}$ and the metric comparison in (6.2) are finally done in one subtract, one test and L_v multiply-accumulate instructions per candidate vector so that the resulting theoretical complexity is approximately

$$\begin{aligned} \mathcal{C}_{S_{\mathrm{FS}}} &= \underbrace{((L_v+1)\cdot(\frac{L_v}{2}+2)+L_v)}_{\text{Instructions per vector}} \cdot \underbrace{N_g \cdot N_{\mathrm{svq}}}_{\substack{\text{Number of}\\ \text{candidates per}\\ \text{vectors}}} \cdot \underbrace{\frac{f_s}{L_v}}_{\substack{\text{Vectors}\\ \text{per second}}} \\ &\approx 549 \text{ GIPS} \end{aligned} \tag{6.3}$$

with GIPS standing for Giga-instructions per second (1GIPS$\hat{=}2^{30}$ instructions per second).

The corresponding highest achievable SNR in dB is defined as $\mathrm{SNR}_{S_{\mathrm{FS}}}|_{\mathrm{dB}}$. In order to get a reference value for the standard configuration from Table 6.1 $\mathrm{SNR}_{S_{\mathrm{FS}}}|_{\mathrm{dB}}$ was determined in very time consuming (offline) simulations based on quasi-stationary signals which have been artificially generated as described in Section 5.2.4.5 (more details on the methodology are described in Section E.1.1 of the *supplement document* [Krü09]). In comparison to this reference value, given the same signal to be quantized and the same codec configuration but another search strategy S_{X} with reduced complexity $\mathcal{C}_{S_{\mathrm{X}}} < \mathcal{C}_{S_{\mathrm{FS}}}$, naturally the achievable SNR in dB is equal to or lower than this reference value, $\mathrm{SNR}_{S_{\mathrm{X}}}|_{\mathrm{dB}} \leq \mathrm{SNR}_{S_{\mathrm{FS}}}|_{\mathrm{dB}}$. Accordingly, a logarithmic quality loss measure is defined as

$$q_{S_{\mathrm{X}}}|_{\mathrm{dB}} = \mathrm{SNR}_{S_{\mathrm{FS}}}|_{\mathrm{dB}} - \mathrm{SNR}_{S_{\mathrm{X}}}|_{\mathrm{dB}} \geq 0. \tag{6.4}$$

6.1.3 Complexity Reduction Methods

The *gain-shape decomposition* (strategy S_{GS}) was already developed in Section 5.1.8 and is a method for the reduction of the computational complexity of the SCELP encoder. It is applicable for LSVQ and CELP coding in general. In contrast to this, the other techniques, the *pre-selection* (strategy S_{PS}), the *efficient metric computation* (strategy S_{MC}) and the *candidate exclusion* (strategy S_{CE}) exploit the properties of the Apple Peeling spherical code described in Section 4.4.3 and are in detail explained in Section E.1 of the *supplement document* [Krü09]. Relying on the knowledge of signal $\mathbf{d}''$ and the specification of the weighted vector search (5.24), the proposed techniques can be employed only in combination with the modified CELP approach from Section 5.1.7.1.
The possible reduction of complexity together with the corresponding loss of quantization performance (6.4) is summarized for the listed strategies in Table 6.2. Details on the presented values are given in Section E.1 of the *supplement document* [Krü09]. The results are based on measurements for quasi-stationary signals which have been artificially generated as described in Section 5.2.4.5 and Section E.1 of the *supplement document* [Krü09]. Since it was observed that the quantization performance loss compared to the full search approach depends on the characteristic of the signal to be quantized, an interval of loss measures is provided: The minimum quantization performance loss from Table 6.2 has been measured for uncorrelated signals whereas the maximum loss occured for highly correlated signals. From left to right in Table 6.2, more and more of the methods for complexity reduction are activated to furthermore reduce the complexity. At the same time, the loss of quantization performance increases, only the efficient metric computation is lossless. In the end, with all methods active, the theoretical complexity is as low as 4.5 MIPS. Compared to the full search approach, this means that a complexity reduction of factor

$$\frac{\mathcal{C}_{S_{\mathrm{FS}}}}{\mathcal{C}_{S_{\text{CE-S}}}} = 122222 \approx 1.2 \cdot 10^5 \tag{6.5}$$

Strategy S_{X}	S_{FS}	S_{GS}	$S_{\mathrm{GS}}+S_{\mathrm{PS}}$	S_{GS} + $S_{\mathrm{PS}}+S_{\mathrm{MC}}$	$S_{\mathrm{GS}}+S_{\mathrm{PS}}$ $+S_{\mathrm{MC}}+S_{\mathrm{CE}}$
Estimate of theoretical complexity $\mathcal{C}_{S_{\mathrm{X}}}$	549 GIPS	19 GIPS	185 MIPS	35 MIPS	4.5 MIPS
min. loss $\sum q_{S_{\mathrm{X}}}\|_{\mathrm{dB}}$	0 dB	0 dB	0 dB	0 dB	0.5 dB
max. loss $\sum q_{S_{\mathrm{X}}}\|_{\mathrm{dB}}$	0 dB	1 dB	2 dB	2 dB	2.5 dB

Table 6.2: Summary of performance loss due to different codevector search strategies. The minimum loss is the estimate for uncorrelated signals, the maximum for highly correlated signals. GIPS $\hat{=}$ Giga-instructions Per Second, MIPS $\hat{=}$ Million Instructions Per Second.

can be achieved while introducing only a quantization performance loss of

$$0.5 \text{ dB} \lesssim q_{S_{\mathrm{SCELP}}} \lesssim 2.5 \text{ dB}. \tag{6.6}$$

Note that the techniques for reduced complexity codevector searches can principally be combined with both, the joint (Section 5.1.8.1) and the sequential (Section 5.1.8.2) approach for the gain-shape decomposition. However, the described minimum computational complexity and quality loss can only be achieved if combined with the sequential approach.
The overall performance loss for correlated sources may seem to be high enough to motivate the search for different approaches for quantization. From another point of view, however, an SNR loss of 2.5 dB is not very significant in cases where the achieved audio coding quality already is very high as for highly correlated sources. Most important for the overall quality of a codec is in general the behavior for signals which are hard to be coded. Since the SNR is low for signals which have a low prediction gain (uncorrelated sources), the performance loss of 0.5 dB due to the proposed complexity reduction methods does not have a relevant impact on the overall audio quality achieved by the SCELP codec.

6.2 W-SCELP: Extending SCELP by *Warped Linear Prediction*

So far, aspects of human perception were only marginally considered. In perceptual audio coding, however, these aspects are of very high importance [PS00]. The noise

shaping due to the error weighting filter from (5.14) or (5.26) is commonly the basis for perceptual masking in speech coding and has been intensively investigated in the literature. It was shown in the previous chapter, however, that the conventional error weighting filter (5.14) at the same time can be considered as a suboptimal technique for reverse waterfilling (Section 5.2.6.2). The new optimization criterion from Section 5.2.6 leads to a better approximation of the reverse waterfilling behavior in LPC but, nevertheless, can be combined well with the conventional noise shaping. In this context, another degree of freedom is available in the SCELP and the W-SCELP codecs to control reverse waterfilling and perceptual noise masking independently.

In conventional LPC, the approximation of the spectral envelope of the input signal is based on a uniform resolution of the frequency scale. Considering the perceptual properties of human hearing, a uniform resolution is known to be inferior compared to a non-uniform resolution of the frequency scale [ZF99]. Therefore, the approach of (frequency) warped linear prediction (WLP) shall in the following be introduced and adapted for the purpose of CELP encoding to contribute for improved perceptual noise masking. The principle of warped signal processing and WLP was at first presented in [Str80]. Realization aspects related to WLP were, e.g., presented in [HÜ98] and [HÜ00]. An example for WLP in source coding is explained in [HL99], and in [HL01], a formal comparison between warped and conventional linear prediction points out the advantages of WLP in the context of audio coding. The application of WLP for audio coding is mostly straightforward but requires a few important modifications which are described in [KV07a]. Only the principle shall be briefly discussed here.

The weighted vector search involving only the truncated impulse response of the combined weighting filter in the modified structure of the CELP encoder from Section 5.1.7.1 allows to combine WLP with the SCELP techniques without significant additional computational effort. This combination is denoted as the Warped Spherical Code-Excited Linear Prediction (W-SCELP) codec.

6.2.1 Principle of Warped Linear Prediction (WLP)

A non-uniform resolution of the frequency scale can be achieved by frequency warping which, given a system function in the z-domain, can be realized by replacing all unit delay elements by the allpass filter $A^w(z)$,

$$z^{-1} \rightarrow A^w(z) = \frac{z^{-1} - \lambda^w}{1 - \lambda^w \cdot z^{-1}} \quad |\lambda^w| < 1;\ \lambda^w \in \mathbb{R}. \tag{6.7}$$

The employment of this principle in the context of LP is denoted as warped linear prediction (WLP). The non-uniform resolution of the frequency scale is controlled by warping constant λ^w [SA99]: For positive values the spectral resolution is high for lower and low for higher frequencies. A warping coefficient of, e.g., $\lambda^w = 0.57$ was described in [HL01] to best approximate the well known Bark frequency scale [Zwi61] in wideband coding (sample rate $f_s = 16$ kHz).

All important aspects involved in LPC can be easily extended towards warped linear prediction (refer to [KV07a] also),

- the LP analysis based on the autocorrelation method (see Section 5.1.2).
- the LP analysis and synthesis filter.
- the error weighting filter $W(z)$ from (5.26).
- the reverse waterfilling in CELP coding (Section 5.2.6.2).

6.2.2 Implementation Aspects for WLP and Source Coding

Nevertheless, a direct application of the warped LP analysis, synthesis and error weighting filters is prohibitive. Therefore, the following modifications must be introduced in the W-SCELP low delay audio codec:

- Removal of the zero-delay path in the feedback loop of the LP synthesis and the error weighting filter.
- Introduction of the zero-mean property for all WLP filters.
- Spectral tilt compensation for the WLP filters.
- Combination of WLP and CELP Coding.

These aspects are discussed in detail in [KV07a].

6.2.3 Conventional and Warped LP: A qualitative Comparison

A qualitative comparison of conventional LP and WLP is exemplified by Figure 6.1. In that figure, it is shown how the conventional LP synthesis filter ($H_S(z)$) and the warped LP synthesis filter ($H_S^w(z)$) approximate the spectral envelope of an example signal $x(k)$. The (qualitative) squared magnitude spectrum $|X(\Omega)|^2$ for signal $x(k)$ in dB is shown as the gray line, the magnitude spectra $|H_S(\Omega)|^2$ and $|H_S^w(\Omega)|^2$ related to the LP and WLP synthesis filter spectra computed from the corresponding transfer functions with $z = e^{j\cdot\Omega}$ in dB are depicted as the solid and the dotted black line, respectively. The LP order for the approximation of the spectral envelopes is $N_{\text{lpc}} = 10$. All magnitude spectra have been computed based on large DFTs and are shown for the complete range of normalized frequencies $0 \leq \Omega \leq \pi$. The warping factor for the WLP based approximation is $\lambda^w = 0.5$. Clearly, the WLP based approximation is significantly more accurate for lower frequencies than the approximation based on conventional LP.
Comparing the achievable prediction gain as a measure for signal decorrelation, WLP provides only an insignificantly higher value than conventional LPC. Nevertheless, WLP leads to a higher perceived quality since it is better consistent with the properties of human perception than conventional LP. The difference compared to conventional LP is significant especially for audio signals with a sparse spectrum with strong low-frequency components, for example the sound of a flute.

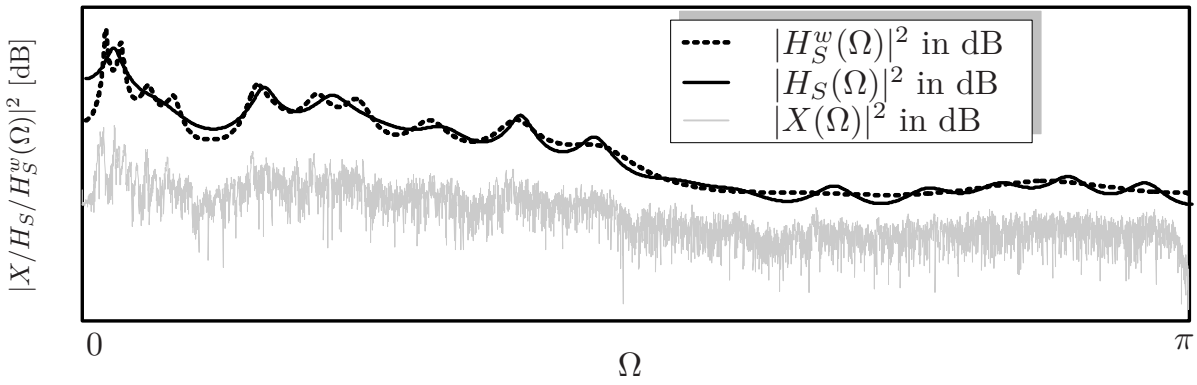


Figure 6.1: Example for the approximation of the spectral envelope of an example signal by means of the LP and WLP synthesis filters.

	W-SCELP Codec		SCELP Codec	
	Encoder	Decoder	Encoder	Decoder
Estimated Complexity in WMOPS	23-28	2-3	20-25	1-2

Table 6.3: Measured overall computational complexity of the W-SCELP and the SCELP low delay audio codec for the standard configuration (Table 6.2). The given values were measured for an implementation in fixed point arithmetic based on the set of fixed point operations specified by the ITU-T in [ITU00].

6.3 Measured Computational Complexity in Practice

Both, the W-SCELP and the SCELP codec were realized in floating and fixed point arithmetic. Since no specific signal model is employed, the floating point version of both codecs is highly scalable and can adapt to various application scenarios and constraints. In contrast to this, only a specific configuration (the standard configuration from Table 6.1) was realized in fixed point arithmetic based on the set of fixed point operations as proposed by the ITU in [ITU00]. This set of operations enables to realize algorithms in digital signal processing in fixed point arithmetic based on a virtual instruction-set-architecture (ISA). In order to represent the effort to realize it on state-of-the-art DSPs, each employed computational operation is weighted to yield the overall computational complexity in Weighted Million Operations Per Second (WMOPS). The measured results are not an exact estimate of the complexity to be expected when porting the W-SCELP codec to a specific DSP platform but more useful for first estimates of the involved complexity than the theoretical complexity discussed earlier. The measured complexity of the W-SCELP and the SCELP codec is listed in Table 6.3. More details describing the realization of the W-SCELP and the SCELP codec in fixed point arithmetic are given in [KSV06].
In order to demonstrate the high efficiency of the SCELP and the W-SCELP codec,

real-time prototypes for both codecs were created by means of the RTProc system [KLEV03], [KV08c], [KJLV09], and [KSE+09].

6.4 Quality Assessments

The overall audio quality of the SCELP and the W-SCELP low delay audio codecs was evaluated in assessments based on objective audio quality measures for a comparison with the ITU-T G.722 audio codec [ITU88b] at the three supported bit rates of 48, 56 and 64 kbit/sec. The ITU-T G.722 codec was chosen as a reference codec because of its algorithmic delay in the magnitude of that of the W-SCELP codec (below 10 ms). Another alternative codec with a comparable algorithmic delay would have been the *Advanced Audio Codec - Ultra Low Delay* (AAC-ULD) [Fra07]. This codec, however, is not freely available and could thus not be considered here.
In order to be comparable to the ITU-T G.722 codec for wideband audio, the SCELP and the W-SCELP codec were operated at a sample rate of $f_s = 16$ kHz and a bit rate of approximately 48 kbit/sec rather than in the standard configuration from Table 6.2 and were configured to achieve an algorithmic delay of 9 ms. A warping factor of $\lambda = 0.46$ was determined in informal listening tests to configure the W-SCELP codec for higher perceived audio quality.
In order to form a complete codec, besides the quantization of the LP residual signal, a dedicated quantizer was developed to encode the LP coefficients. The quantization of LP coefficients in speech coding was intensively investigated in the literature, and is in general possible at very low bit rates [PA93]. Most of the approaches from the literature are based on the transformation of the LP coefficients into *Line Spectral Frequencies* (LSFs), e.g., [Ita75], and intensive training of VQ codebooks for large speech databases. For audio coding, however, due to the different characteristics of audio signals compared to speech, new approaches for the computation and the quantization of LSFs were developed, e.g., a new approach for the quantization of LSFs based on LSVQ and a two-dimensional inter-and intra-frame predictor. These techniques are in detail described in [KSV06] and shall not be discussed here. With the techniques for the transmission of LSFs for audio signals, required bit rates to transmit LP coefficients for linear prediction of order $N_{\mathrm{lpc}} = 10$ of approximately 4 kbit/sec were achieved. Given a bit rate of 48 kbit/sec in the SCELP and the W-SCELP codec, the transmission of the LP coefficients hence requires approximately 10 percent of the overall data rate. The required parameters C_Δ, γ_1 and γ_2 to configure the reverse waterfilling and the noise shaping according to Section 5.1.7.2, respectively, were determined in informal listening tests prior to the actual quality assessment.
For the comparison of the perceptual audio quality, two databases, one for speech and one for audio signals were processed by the SCELP, the W-SCELP, and the ITU-T G.722 audio codec. For the assessment of the codecs for speech signals, the decoder output signals were evaluated by the *wideband perceptual evaluation*

of speech quality measure (WB-PESQ), [ITU05]) which is widely accepted in the speech coding community. The assessment for audio signals was based on the counterpart of the WB-PESQ for audio signals, the *perceptual evaluation of audio quality* measure (PEAQ)[ITU98][Thi00].
Note that the quality was assessed based on monaural speech and audio signals. New hierarchical approaches to extend the SCELP and the W-SCELP towards coding of stereo signals are proposed in [KV08b], [KV08a] and [SKV09]. These aspects, however, shall not be discussed here.

6.4.1 Results for Speech Signals

The measured objective quality of the W-SCELP codec in comparison to that of the ITU-T G.722 codec based on the WB-PESQ measure is summarized in Table 6.4. WB-PESQ based quality measures are specified on a scale from 0 to 5 MOS (*Mean Opinion Score*), and a higher achieved speech quality leads to a higher WB-PESQ value. The performance of the ITU-T G.722 codecs was rated with 4.02, 4.39 and 4.47 MOS for the three codec modes respectively. The SCELP at a data rate of roughly 48 kbit/sec reached a value of 4.4 MOS. The quality of the SCELP codec at 48 kbit/sec can hence be classified as slightly better than that of the ITU-T G.722 codec at 56 kbit/sec for speech signals. The W-SCELP codec is not explicitly listed here since SCELP and W-SCELP codec have the same performance for speech signals.

Codec	G.722 mode 1	**SCELP**	G.722 mode 2	G.722 mode 3
Data rate	$64\frac{\text{kbit}}{\text{sec}}$	$48\frac{\text{kbit}}{\text{sec}}$	$56\frac{\text{kbit}}{\text{sec}}$	$48\frac{\text{kbit}}{\text{sec}}$
WB-PESQ (MOS-LQO)	4.47	4.4	4.39	4.02

Table 6.4: Results from formal quality assessment of the ITU-T G.722 and the SCELP for a data base composed of speech signals. The quality assessment is based on the WB-PESQ measure [ITU05] with MOS-LQO as the objective mean opinion score from listening-only test scenario. The W-SCELP is not explicitly listed here since both, the W-SCELP and the SCELP codec have equal performance for speech signals.

6.4.2 Results for Audio Signals

For audio signals, the difference between the W-SCELP and the SCELP is more significant than for speech signals. Especially for signals with very tonal components in the lower frequency areas, the frequency warping leads to a significant improvement. The results from the evaluation of an audio database are given in Figure 6.2. The audio quality was measured by means of the PEAQ measure based on the implementation from [Kab02]. The PEAQ measurement tool returns values between 0 and -4 to indicate the signal deterioration. The definition of the quality

degradation and corresponding PEAQ values is listed on the left side of Figure 6.2. The measured results for the ITU-T G.722, the SCELP, and the W-SCELP codec are given on the right side of that figure. In addition, also measured results for the well-known MP3 audio codec (implementation within the commercial product *Adobe Audition* [Ado07]) are given which has a delay which is higher than 80 ms[2]. As result, the SCELP and especially the W-SCELP codec outperform the ITU-T

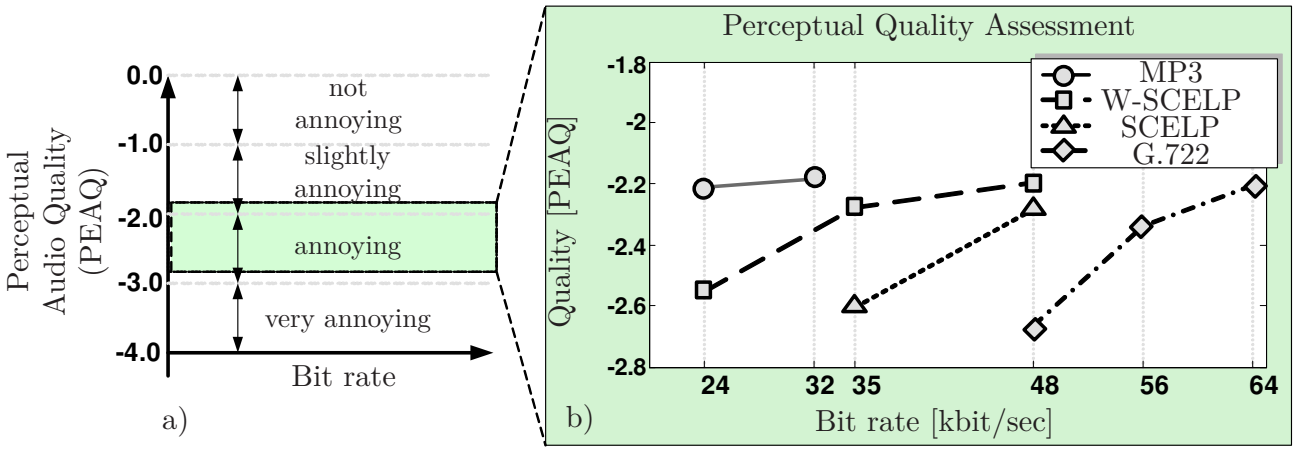


Figure 6.2: Classification of audio quality according to PEAQ measure (a) and measured PEAQ values (based on a data base composed of audio files) for the W-SCELP, the SCELP, the ITU-T G.722 and the MP3 audio codec (b). Delay ITU-T G.722, W-SCELP, SCELP Codec < 10 ms, Delay MP3 Codec > 80 ms.

G.722 codec significantly:

- The quality achieved by the W-SCELP codec at a bit rate of 48 kbit/sec is approximately equal to that of the ITU-T G.722 codec at 64 kbit/sec.
- Even for the lower bit rate of 35 kbit/sec, the W-SCELP achieves an audio quality which is only slightly worse than the ITU-T G.722 codec at 64 kbit/sec.
- An audio quality comparable to that achievable by the ITU-T G.722 codec at a bit rate of 48 kbit/sec can approximately be achieved by the W-SCELP codec at bit rates of 24 kbit/sec.

In comparison to the MP3 codec at 32 kBit/sec, the W-SCELP codec produces almost a comparable audio quality at a bit rate of 35 kBit/sec. In this comparison, however, it has to be considered that the W-SCELP has an algorithmic delay of less than 10 ms which is (at least) by a factor of 8 lower than that of the MP3 codec. Also, the MP3 codec does not code the full audio frequency bandwidth which, even though it degrades the audio quality noticeably, does not lead to an adequate reduction of the PEAQ measure.
As a conclusion of the measurements, for applications such as wireless microphones for live concerts and wireless audio-links for hearing aids, only the SCELP and

[2]The delay of the MP3 codec is not exactly specified but was measured to be at least 80 ms

the W-SCELP deliver a sufficient audio quality **and** fulfill the technical constraints defined in Section 1.

7

Summary

Most systems for the transmission and storage of speech and audio signals are nowadays based on digital technology. For specific applications, however, operation constraints are defined which only analog technology could fulfill. One of the most critical operation constraints is to achieve a low algorithmic delay. Informal listening tests showed that for specific applications, a delay of more than 10 ms is not acceptable. Most standardized audio codecs have a significantly higher delay or were designed for speech only. Consequently, no standardized digital audio codec is currently available which enables an algorithmic delay below 10 ms and at the same time achieves a high perceptual quality for speech and audio signals at a low bit rate and with low computational cost.
The outcome of this thesis are novel techniques for the lossy compression of speech and audio signals applicable for low delay source coding as well as two new low delay audio codecs, the Spherical Code-Excited Linear Prediction (SCELP) codec and its successor, the Warped Spherical Code-Excited Linear Prediction (W-SCELP) codec.

Concept for Low Delay Audio Coding
In order to achieve a low algorithmic delay, the main concept followed in this thesis is combined linear prediction and vector quantization which is well-known from state-of-the-art speech codecs. Most standardized codecs relying on this principle, however, have an algorithmic delay of more than 20 ms and/or were designed for speech only, e.g., the Adaptive Multirate speech codec [ETS00]). Therefore, fundamental modifications of the concepts known from speech coding are essential to achieve an algorithmic delay below 10 ms and a low computational complexity as well as to enable high perceived coding quality at low bit rates also for audio signals:

- No speech specific components as known from speech coding such as the long term prediction (LTP) to model the speaker instantaneous pitch period are employed.

- In return, a novel type of vector quantizer, the Logarithmic Spherical Vector Quantizer (LSVQ), is employed to achieve a higher perceived quality at marginally increased bit rates compared to speech coding.

- A new model for combined linear prediction and quantization with quantization error feedback valid for high as well as low bit rates is proposed for stability and performance analyses and the optimization of codec parameters.

- Based on the new model a novel optimization criterion for the LP analysis at low bit rates is derived to account for the *reverse waterfilling* (RW) principle known from rate distortion theory (RDT) due to the properties of audio signals and the low algorithmic delay of the codec.

- Warped linear prediction (WLP) better exploits the properties of human perception than conventional linear prediction due to a non-uniform resolution of the frequency scale and hence achieves a higher perceptual audio quality especially for audio signals.

The development of the new techniques and especially the two approaches for low delay audio coding are based on fundamental novel theoretical results on quantization and linear predictive coding presented in this thesis. Practical relevance is retained, however, since all theoretical investigations are accompanied by novel practically relevant concepts with special focus on low computational complexity.

Logarithmic Spherical Vector Quantization

Logarithmic Spherical Vector Quantization (LSVQ) is the direct consequence of the (asymptotic) high rate vector quantization theory to approximate the optimal codevector density for stationary signals with multivariate identical independent Gaussian distribution. In contrast to speech coding where normalized LP residual signals are known to be approximately Gaussian distributed, in audio coding, the assumption of a Gaussian distribution is useful to define a worst case scenario to guarantee a minimum audio quality for all types of signals.
In LSVQ, each input signal vector is decomposed into a gain factor and a shape vector. Both components are quantized by means of logarithmic scalar quantization and Spherical Vector Quantization (SVQ), respectively. In the present thesis, novel **qualitative** results are derived for high rate assumptions to show that the signal-to-noise ratio (SNR) achieved by LSVQ is independent of the distribution of the input signal in a wide dynamic range. This result is consistent with the quantizer design target to deliver a minimum audio quality for all types of signals. The definition of an **idealized** SVQ is the basis for the derivation of a **quantitative** expression for an upper SNR bound for LSVQ. This expression allows to determine the optimal allocation of bits to the quantizers for the gain and the shape component given an overall LSVQ bit budget. In addition, it exhibits that a high vector dimension for LSVQ is desirable to maximize the quantization performance, but the performance gain due to an increased vector dimension decreases for higher dimensions and may no longer justify the involved additional computational effort in practice. LSVQ is

shown to be asymptotically optimal for stationary signals with multivariate identical independent Gaussian distribution.
In addition to the theoretical investigations, three practical concepts for LSVQ are investigated. Due to the development of new efficient algorithms for nearest neighbor quantization, all three concepts can be realized with low computational complexity and therefore are of high practical relevance. Measurements of the quantization SNR and comparison to the theoretical results exhibit that all three approaches are well suitable for highly efficient quantization.

Combined Linear Prediction and LSVQ

The combination of quantization and linear prediction (LP) in **closed-loop** manner is known to enable to transform correlation into an increased overall coding SNR. A (generalized) closed-loop combination of a **scalar quantizer** and linear prediction can be realized in the encoder by feeding back a filtered version of the introduced quantization noise. Employing **vector quantization**, the decomposition of the input signal into vectors is contradictory to a realization of closed-loop quantization as a sample-by-sample linear filtering of the quantization error. Therefore, combined closed-loop LSVQ and LP is realized based on the well-known Code-Excited Linear Prediction (CELP) approach.
Theoretical investigations of linear predictive coding (LPC) described in the literature are in general based on high rate assumptions and are therefore valid only for high bit rates. In this thesis, a new theory of LPC is developed that is also valid for lower bit rates. The new theory is based on a novel scalar *noise production and propagation model* which also considers the interaction between the feedback of the quantization error and the quantizer.
The new theory confirms the results from conventional theory of LPC for high bit rates but exhibits that the overall closed-loop LPC encoder can become unstable for specific signals. Also, it is shown that the feedback of quantization noise degrades the achievable quantization performance for low bit rates compared to the results as predicted by the conventional theory. The derived theoretical results are confirmed by measurements of SNRs for artificial stationary signals.
Based on the theory, a new optimization criterion for the computation of the LP and the error weighting filter coefficients for lower bit rates is developed. With respect to this optimization criterion a two-step procedure is proposed which is well applicable in practice. The new optimization criterion and the two-step computation rule turn out to be an approximation of the reverse waterfilling procedure which is known from rate distortion theory but which has never been explicitly described for LPC in the literature. The presented new theory is of high relevance for low delay audio coding as well as for a deeper fundamental understanding of LPC in general.

The SCELP and the W-SCELP Codec

In the last part of the thesis, the SCELP low delay audio codec is developed, based on the combination of LSVQ with linear prediction according to the CELP principle. Practical aspects such as the computational complexity and the regard of the

principles related to the human perception are of special interest. It turns out that, if combined with linear prediction, particularly one among the three approaches proposed for LSVQ enables to achieve a low computational complexity and a high quantization performance at the same time. Compared to the complexity related to the original definition of the CELP principle, novel optimization strategies allow to reduce the complexity by a factor of approximately 10^5 while the quantization performance is only marginally degraded.
A higher perceptual coding quality especially for audio signals is achieved by the W-SCELP codec which is based on the extension of the SCELP codec by warped linear prediction (WLP). WLP is a technique to enable a non-uniform resolution of the frequency scale which is beneficial for coding compared to a uniform resolution since it better matches the properties of human perception.
Since the SCELP and the W-SCELP low delay audio codec do not rely on any signal model, the overall concept is highly scalable and can be adapted to various application scenarios. Measurements based on fixed point implementations of the SCELP and the W-SCELP exhibit that a moderate computational complexity can be achieved in practice.

Objective quality assessments for speech and audio signals based on the well-known wideband perceptual speech quality (WB-PESQ) and the perceptual audio quality (PEAQ) measures show that both, the SCELP and the W-SCELP codec, significantly outperform standardized codecs with a comparable delay and bit rate, e.g., the ITU-T G.722 codec, in terms of a higher subjective quality for speech and particularly audio signals.

A

Derivations of Selected Equations

In Chapter 4, derivations are made to compute distortions and signal-to-noise ratios (SNRs) in the context of the analysis of Logarithmic Spherical Vector Quantization (LSVQ). In this Appendix, selected equations are derived and explained more in detail.

A.1 Additional Explanation for Equation (4.47)

In equation (4.47), an expression to compute the angular radius $\beta_{\max}$ is given for the assumption of high bit rates. This expression can be derived as follows:
It was shown in (4.35) that the area content of a cap quantization cell can be computed from $\beta_{\max}$ as

$$S_{C_{\tilde{\mathbf{c}}}}^{(II)}(r') = V_{\mathcal{S}_{L_v-1}}^{(1.0)} \cdot (L_v - 1) \cdot \int_0^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta. \tag{A.1}$$

This term can be simplified for the assumption of high bit rates due to the following approximations (4.44-4.46)

$$r' \approx 1.0 \tag{A.2}$$

$$r' \cdot \sin(\beta) \approx \sin(\beta) \approx \beta \tag{A.3}$$

$$1 - r' \cdot \cos(\beta) \approx 1 - r' \tag{A.4}$$

to find

$$S_{C_{\tilde{\mathbf{c}}}}^{(II)}(1.0) \approx V_{\mathcal{S}_{L_v-1}}^{(1.0)} \cdot (L_v - 1) \cdot \int_0^{\beta_{\max}} \beta^{(L_v-2)} \cdot \mathrm{d}\beta \tag{A.5}$$

$$= V_{\mathcal{S}_{L_v-1}}^{(1.0)} \cdot (L_v - 1) \cdot \frac{\beta_{\max}^{L_v-1}}{L_v - 1}. \tag{A.6}$$

Also, it was motivated that, in order to compute a lower bound for the distortion, the complete surface area is assumed to be covered by cap quantization cells (4.32):

$$S_{\mathcal{S}_{L_v}}^{(1.0)} = N_{\text{svq}} \cdot S_{C_{\tilde{\mathbf{c}}}}^{(II)}(1.0), \tag{A.7}$$

By substituting (A.6) in (A.7), with (4.5) and by writing $V_{\mathcal{S}_{L_v-1}}^{(1.0)}$ according to (4.3) yields

$$N_{\text{svq}} \cdot V_{S_{L_v-1}}^{(1.0)} \cdot \beta_{\max}^{L_v-1} = V_{S_{L_v}}^{(1.0)} \cdot L_v \tag{A.8}$$

$$N_{\text{svq}} \cdot \frac{\pi^{\frac{L_v-1}{2}}}{\Gamma(\frac{L_v+1}{2})} \cdot \beta_{\max}^{L_v-1} = \frac{2 \cdot \pi^{\frac{L_v}{2}}}{\Gamma(\frac{L_v}{2})} \tag{A.9}$$

from which (4.47) can be computed.

A.2 Additional Explanation for Equation (4.48)

Equation (4.48) is an expression for the approximation of the quantization distortion for high bit rates. It is based on the exact solution given in (4.41) as

$$D^{*(II)}_{\text{lsvq}} = \frac{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int_{0}^{\beta_{\max}} ((r' \cdot \sin(\beta))^2 + (1 - r' \cdot \cos(\beta))^2)(r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'}{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \int_{0}^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot r' \cdot \mathrm{d}\beta \cdot \mathrm{d}r'} \tag{A.10}$$

Again the approximations (4.44-4.46) are used for the assumption of high bit rates,

$$r' \approx 1.0 \tag{A.11}$$

$$r' \cdot \sin(\beta) \approx \sin(\beta) \approx \beta \tag{A.12}$$

$$1 - r' \cdot \cos(\beta) \approx 1 - r'. \tag{A.13}$$

With these, the integral can be approximated as

$$D^{*(II)}_{\text{lsvq}} \approx \frac{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \left(\int_{0}^{\beta_{\max}} \beta^2 \cdot (\beta)^{(L_v-2)} \mathrm{d}\beta + \int_{0}^{\beta_{\max}} (1 - r') \cdot \beta^{(L_v-2)} \mathrm{d}\beta \right) \mathrm{d}r'}{\int_{1-\frac{\Delta g}{2}}^{1+\frac{\Delta g}{2}} \left(\int_{0}^{\beta_{\max}} (r' \cdot \sin(\beta))^{(L_v-2)} \cdot \mathrm{d}\beta \right) \cdot \mathrm{d}r'}. \tag{A.14}$$

By substituting $r'' = 1 - r'$ in the second part of the numerator, this equation can be rewritten as

$$D^{*(II)}_{\text{lsvq}} \approx \frac{\int_0^{\beta_{\max}} \beta^{L_v} \mathrm{d}\beta \cdot \Delta_g}{\int_0^{\beta_{\max}} \beta^{(L_v-2)} \mathrm{d}\beta \cdot \Delta_g} + \frac{\int_{\frac{\Delta_g}{2}}^{-\frac{\Delta_g}{2}} \left(\int_0^{\beta_{\max}} \cdot \beta^{(L_v-2)} \mathrm{d}\beta \right) \cdot r''^2 (-\mathrm{d}r'')}{\int_0^{\beta_{\max}} \beta^{(L_v-2)} \mathrm{d}\beta \cdot \Delta_g} \tag{A.15}$$

$$= \frac{\int_0^{\beta_{\max}} \beta^{L_v} \mathrm{d}\beta}{\int_0^{\beta_{\max}} \beta^{(L_v-2)} \mathrm{d}\beta} + \frac{\Delta_g^2}{12}. \tag{A.16}$$

A.3 Additional Explanation for Equations (4.53-4.55)

In Chapter 4.2.4.4, the optimal allocation of bits for the quantizers of the gain and the shape component, respectively, is given. Starting point for the computation of the optimal bit allocation is the auxiliary function in (4.52),

$$\chi = \frac{C_{\text{svq}}}{N_{\text{svq}}^{\frac{2}{L_v-1}}} + \frac{C_g}{N_g^2} + \lambda \cdot (N_g \cdot N_{\text{svq}} - N_{\text{lsvq}}). \tag{A.17}$$

For the optimization, the derivative of the auxiliary function with respect to N_g and N_{svq} is computed as

$$\frac{\partial \chi}{\partial N_{\text{svq}}} = \lambda \cdot N_g - \frac{2}{L_v - 1} \cdot \frac{C_{\text{svq}}}{N_{\text{svq}}^{\frac{2}{L_v-1}-1}} \tag{A.18}$$

$$\frac{\partial \chi}{\partial N_g} = \lambda \cdot N_{\text{svq}} - 2 \cdot \frac{C_g}{N_g^3} \tag{A.19}$$

In order to find a minimum, the partial derivatives must be set equal to zero. N_{svq} can be computed from (A.19) and is substituted in (A.18) to yield

$$\lambda = \frac{2 \cdot C_g}{N_{\text{svq}} \cdot N_g^3}. \tag{A.20}$$

(A.20) is substituted in (A.18), and the resulting expression is set to zero yielding

$$\frac{2 \cdot C_g}{N_{\text{svq}} \cdot N_g^2} - \frac{2}{L_v - 1} \cdot \frac{C_{\text{svq}}}{N_{\text{svq}}^{\frac{2}{L_v-1}-1}} \stackrel{!}{=} 0 \tag{A.21}$$

With (4.49) and the first part of (4.48), (A.21) can be transformed into the intermediate result given in (4.53),

$$D_g = \frac{D^{*(III)}_{\text{svq}}}{L_v - 1}. \tag{A.22}$$

The definition of the overall bit budget in (4.12) can be modified to find

$$N_g = \frac{N_{\text{lsvq}}}{N_{\text{svq}}} \tag{A.23}$$

which is substituted in (A.21):

$$\frac{C_{\text{svq}} \cdot N_{\text{svq}}^{\frac{-2}{L_v-1}}}{L_v - 1} = C_g \cdot \frac{N_{\text{svq}}^2}{N_{\text{lsvq}}^2}. \tag{A.24}$$

This can be transformed into (4.54),

$$N_{\text{svq}} = \left(\frac{1}{L_v - 1} \cdot \frac{C_{\text{svq}}}{C_g} \right)^{\frac{L_v-1}{2 \cdot L_v}} \cdot N_{\text{lsvq}}^{\frac{L_v-1}{L_v}}. \tag{A.25}$$

The computation of N_g as given in (4.55) can be derived analogously.

A.4 Additional Explanation for Equation (4.56)

Due to the assumption of high bit rates and hence $E\{\| \mathbf{x} \|^2\} \approx 1$, the SNR can be computed from (4.56) as

$$\text{SNR}^{(III)}_{\text{lsvq}} = \frac{1}{D^{*(III)}_{\text{lsvq}}} \tag{A.26}$$

In that equation, the optimal bit allocation from (4.54) and (4.55) is substituted to yield

$$\text{SNR}^{(III)}_{\text{lsvq}} = \frac{1}{\frac{C_{\text{svq}}}{N_{\text{svq}}^{\frac{2}{L_v-1}}} + \frac{C_g}{N_g^2}} \tag{A.27}$$

$$= \left(\frac{C_{\text{svq}}}{\left(\frac{1}{L_v-1} \cdot \frac{C_{\text{svq}}}{C_g} \right)^{\frac{1}{L_v}} \cdot N_{\text{lsvq}}^{\frac{2}{L_v}}} + \frac{C_g}{\left((L_v - 1) \cdot \frac{C_g}{C_{\text{svq}}} \right)^{\frac{L_v-1}{L_v}} \cdot N_{\text{lsvq}}^{\frac{2}{L_v}}} \right)^{-1} \tag{A.28}$$

$$= \left(C_g^{\frac{1}{L_v}} \cdot C_{\text{svq}}^{\frac{L_v-1}{L_v}} \cdot N_{\text{lsvq}}^{\frac{-2}{L_v}} \cdot \underbrace{(L_v - 1)^{\frac{1}{L_v}} \cdot \left(1 + \frac{1}{L_v - 1}\right)}_{= \frac{L_v}{(L_v-1)^{\frac{L_v-1}{L_v}}}} \right)^{-1} \tag{A.29}$$

B

Reproduction of the Presented Results

Most of the theoretical results presented for quantization are in general based on the assumption of stationary signals. In that context, the degree and characteristic of correlation plays a very important role. In order to compare theory and practice, measurements with real audio signals are not well suited since audio signals are in general instationary and consist of signal segments, each with a completely different type of correlation.
For this reason, all theoretical investigations were based on the assumption of AR processes with sets of AR filter coefficients as the origin of all signals to be processed, and all measurements were based on artificial signals produced as the output of the realization of AR processes with the same sets of AR filter coefficients in this thesis. For all investigations with respect to rate distortion theory, assumptions were made about the Eigenvalues of autocovariance matrices. In that context, both, the autocovariance matrices and the sets of AR filter coefficients are parameters to setup signal correlation and therefore equivalent representations.
Due to this very formal approach to produce measurement results, most of the results (except for the quality assessment in Chapter 6) are highly reproducible. The information required for a reproduction is given in this chapter:

- In the first part, the sets of AR filter coefficients for the investigations in Section 5.2.4 and the *supplement document* [Krü09] are listed together with the parameters derived in the context of the noise production and propagation model.

- In the second part, three sets of Eigenvalues involved in the computation of the rate distortion function for different correlated Gaussian sources (Section 2.3.4) are listed.

B.1 Example Sets of AR Filter Coefficients

All computations and measurements of SNRs in Section 5.2.4 and Section E.1 of the *supplement document* [Krü09] are based on the assumption of the signal model shown in the *signal generation* block in Figure 5.7 in which a Gaussian distributed noise source is filtered in an auto-regressive (AR) all-pole filter. The sets of AR filter coefficients are denoted as $\mathbf{a}_{\mathrm{ar},1}$ and $\mathbf{a}_{\mathrm{ar},2}$ in the thesis and listed in Table

AR Filter coefficients	Set $\mathbf{a}_{\mathrm{ar},1}$	Set $\mathbf{a}_{\mathrm{ar},2}$
$a_{\mathrm{ar},0}$	1.0	1.0
$a_{\mathrm{ar},1}$	-2.58425	-1.95498
$a_{\mathrm{ar},2}$	2.95464	1.2564
$a_{\mathrm{ar},3}$	-2.08111	-0.344841
$a_{\mathrm{ar},4}$	1.23315	0.169368
$a_{\mathrm{ar},5}$	-0.920031	-0.340832
$a_{\mathrm{ar},6}$	0.969564	0.00681053
$a_{\mathrm{ar},7}$	-1.22493	0.193209
$a_{\mathrm{ar},8}$	1.51283	0.0923117
$a_{\mathrm{ar},9}$	-1.76435	-0.156275
$a_{\mathrm{ar},10}$	1.72109	0.0103157
$a_{\mathrm{ar},11}$	-1.14735	0.321698
$a_{\mathrm{ar},12}$	0.471528	-0.163375
$a_{\mathrm{ar},13}$	-0.199035	-0.222967
$a_{\mathrm{ar},14}$	-0.18829	0.191995
$a_{\mathrm{ar},15}$	0.667626	0.150169
$a_{\mathrm{ar},16}$	-0.514674	-0.291168
$a_{\mathrm{ar},17}$	0.0287184	-0.0177935
$a_{\mathrm{ar},18}$	0.0887971	0.112793
Prediction Gain $G_{d_0,x}\,\vert_{\mathrm{dB}}$ from (5.48)	19.62 dB	31.5 dB
Feedback Gain $G_{\Delta,e_f}\,\vert_{\mathrm{dB}}$ from (5.48)	15.46 dB	7.8 dB

Table B.1: Two sets of AR filter coefficients labeled as $\mathbf{a}_{\mathrm{ar},1}$ and $\mathbf{a}_{\mathrm{ar},2}$ as examples for theoretical investigations and the generation of artificial signals for measurements. The AR filter order is $N_{\mathrm{ar}} = 18$. In addition to the coefficients, the *prediction* and the *feedback gain* as introduced in the context of the model for LPC and computed according to (5.48) are shown.

B.1. The auto-regressive (AR) all-pole filter order is $N_{\mathrm{ar}} = 18$. In addition, the *prediction* and the *feedback gain* as introduced in (5.48) are given.
In order to represent different types of **real** audio signals, the listed filter coefficients were derived from LP coefficients which were computed by means of an LP analysis for short segments of **real** audio signals. The two listed sets were chosen since they represent common types of signals and are suitable to well demonstrate relevant aspects in theoretical and practical investigations.
The spectral envelopes related to the sets of AR filter coefficients is presented in Figure B.1 as the squared magnitude spectrum $|\ H_{\mathrm{ar}}(\Omega)\ |^2$ computed from the system function

$$H_{\mathrm{ar}}(z) = \frac{1}{H_0(z)} = \frac{1}{\sum_{i=0}^{N_{\mathrm{ar}}} z^{-i} \cdot a_{\mathrm{ar},i}} \tag{B.1}$$

with the coefficients from Table B.1 and $z = e^{j\Omega}$ (FFT size of $L_{\mathrm{fft}} = 8192$).

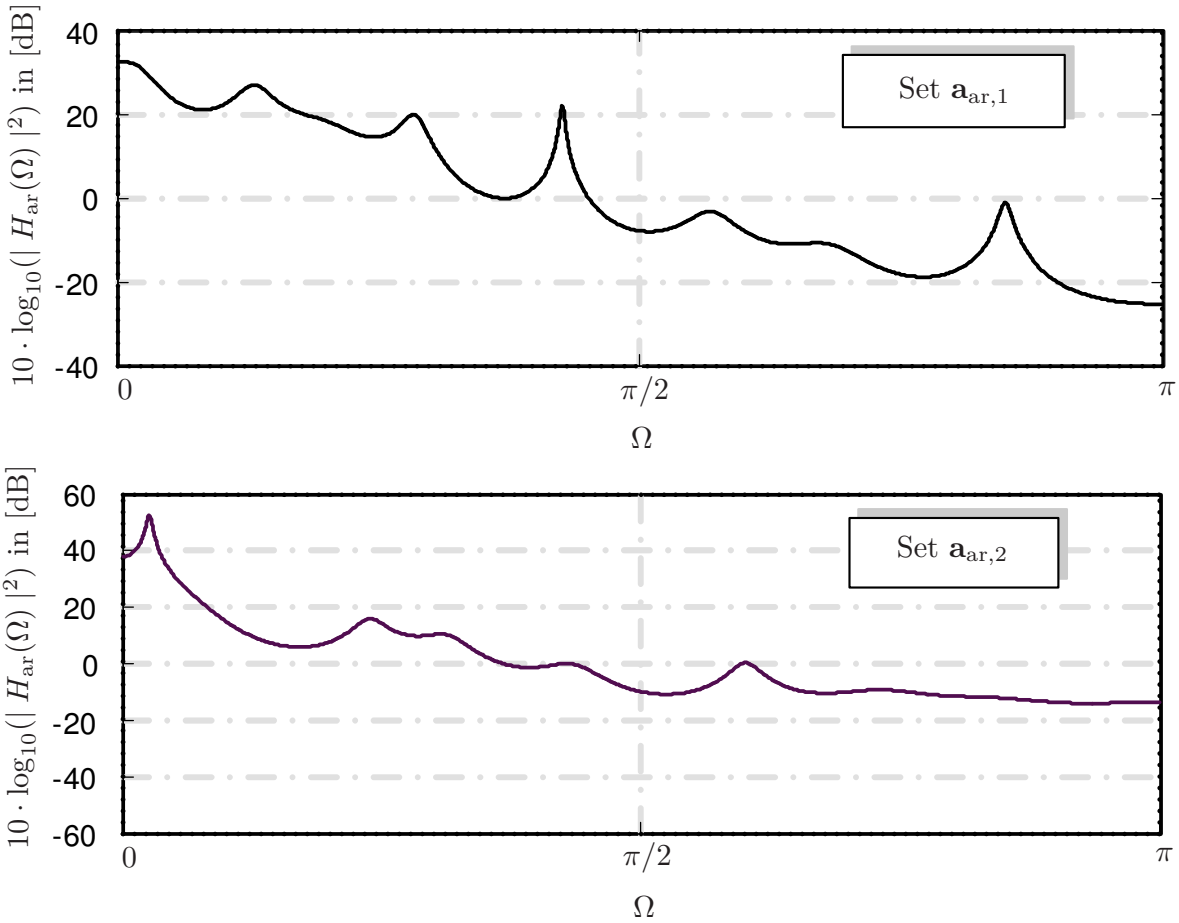


Figure B.1: Logarithmic squared magnitude spectrum $|\ H_{\mathrm{ar}}(\Omega)\ |^2$ in dB computed from system function $H_{\mathrm{ar}}(z)$ for both sets of AR filter coefficients, $\mathbf{a}_{\mathrm{ar},1}$ and $\mathbf{a}_{\mathrm{ar},2}$ from Table B.1.

B.2 Example Eigenvalue Matrices for Rate-Distortion Plots

In Chapter 3, plots of the rate distortion functions related to correlated Gaussian sources with different covariance matrices (and Eigenvalues) were presented. In the following the corresponding three matrices composed of the Eigenvalues are listed to enable a reproduction of the curves in Figure 2.3.4:

$$\mathbf{\Lambda}_{\mathbf{X},a} = \begin{bmatrix} 15.0000 & 0 & 0 & 0 \\ 0 & 0.0008 & 0 & 0 \\ 0 & 0 & 21.0381 & 0 \\ 0 & 0 & 0 & 3.9611 \end{bmatrix} \tag{B.2}$$

$$\mathbf{\Lambda}_{\mathbf{X},b} = \begin{bmatrix} 15.0000 & 0 & 0 & 0 \\ 0 & 0.0100 & 0 & 0 \\ 0 & 0 & 24.7203 & 0 \\ 0 & 0 & 0 & 0.2697 \end{bmatrix} \tag{B.3}$$

$$\mathbf{\Lambda}_{\mathbf{X},c} = \begin{bmatrix} 0.2000 & 0 & 0 & 0 \\ 0 & 0.2000 & 0 & 0 \\ 0 & 0 & 38.9583 & 0 \\ 0 & 0 & 0 & 0.6417 \end{bmatrix} \tag{B.4}$$

C

Deutschsprachige Zusammenfassung

Verfahren zur Übertragung und Speicherung von Sprach- und Audiodaten basieren heutzutage zumeist auf Techniken der digitalen Signalverarbeitung. Insbesondere Technologien zur kompakten Darstellung von digital vorliegenden Signalen mithilfe von effizienten Quellcodierverfahren (Sprach- und Audiocodecs) spielen dabei eine wichtige Rolle.

Einige Anwendungen unterliegen jedoch vorgegebenen Randbedingungen, die bislang nur durch Analogtechnik zu realisieren waren. Beim Einsatz von Drahtlos-Mikrofonen für Live-Konzerte zum Beispiel ist eine geringe Verzögerung der Signalübertragung notwendig, die weniger als 10 ms betragen sollte. Hiervon abweichende, höhere Verzögerungen können unter anderem dazu führen, dass der Nutzer eines Drahtlos-Mikrofons durch die verzögerte Wiedergabe seiner eigenen Stimme gestört wird oder es zu unerwünschten Effekten durch die Überlagerung des verzögerten und des nicht verzögerten Signals kommt. Neben einer geringen Verzögerung stellen eine hohe (wahrnehmbare) Qualität für Sprach- und Audiosignale bei niedrigen Bitraten, Robustheit bei Bitfehlern während der Übertragung sowie eine geringe Rechenkomplexität weitere, wichtige Anforderungen in der Praxis dar. Diese sind zum Teil jedoch widersprüchlich.

Die Standardisierung von Verfahren zur digitalen Quellcodierung von Sprachsignalen auf der einen und Audiosignalen auf der anderen Seite wurde in der Vergangenheit von unterschiedlichen Interessensgruppen verfolgt. Grund hierfür ist die Definition verschiedener Anwendungsziele und damit verbundene sich grundlegend unterscheidende praktische Randbedingungen. Entsprechend müssen zwei Familien von Quellcodierverfahren unterschieden werden:

Verfahren zur Sprachcodierung dienen in der Regel der Codierung von Schmalbandsprache (300-3400 Hz Audiobandbreite) oder Breitbandsprache (50-7000 Hz Audiobandbreite) für die mobile Sprachkommunikation. Relativ geringe Verzögerungen sind hierbei in erster Linie für das Erreichen einer hohen Komunikationsqualität im Duplex-Modus wichtig. Darüber hinaus spielt eine niedrige Rechenkomplexität in der Regel eine entscheidende Rolle, um die Betriebsdauer von Batterie betriebenen mobilen Kommunikationsendgeräten zu maximieren. Bei den am weitesten verbreiteten Standards zur Sprachcodierung liegt die algorithmische Verzögerung in der Regel in einer Größenordnung von 20-30 ms, was für die geforderten Randbedingungen (z.B. für Drahtlos-Mikrofone) zu hoch ist. Einige spezielle Codecs weisen eine niedrigere Verzögerung auf, die im Bereich von unter 10 ms liegen kann. Durch die Ausnutzung der Eigenschaften von Sprache erreichen Verfahren zur Sprachcodierung jedoch für Audiosignale eine nur unzureichende Qualität.

Verfahren zur Audiocodierung dienen vorwiegend der Speicherung von Audiodaten (z.B. zwecks Archivierung) und zielen darauf ab, möglichst hohe bzw. transparente Qualität zu erreichen. Geringe algorithmische Verzögerung und Rechenkomplexität spielten bei der Entwicklung bislang eine eher untergeordnete Rolle. Aus diesem Grunde ist die Verzögerung der gängigen Verfahren zur Audiocodierung deutlich höher als bei der Sprachcodierung. Darüber hinaus sind Verfahren zur Audiocodierung gegenüber Bitfehlern in der Regel deutlich weniger robust als Sprachcodecs und deswegen für die Übertragung über gestörte Kanäle ungeeignet.

Existierende standardisierte Verfahren der digitalen verlustbehafteten Quellencodierung können die eingangs geforderten Randbedingungen derzeit nur unzureichend erfüllen. Gegenstand dieser Dissertation ist deshalb die Entwicklung neuartiger Verfahren zur effizienten verlustbehafteten Codierung von digitalen Sprach- **und** Audiosignalen bei gleichzeitig geringer Latenz. Konkrete praktische Anwendung finden die entwickelten Verfahren in zwei neuen Audiocodecs, dem *Spherical Code-Excited Linear Prediction* (SCELP) und dem *Warped Spherical Code-Excited Linear Prediction* (W-SCELP) Audiocodec mit einer Latenz von jeweils unter 10 ms und einer moderaten Rechenkomplexität.

Zugrunde liegendes Konzept zur Sprach-Audiocodierung

Um eine niedrige algorithmische Verzögerung bei der Audiocodierung zu erreichen, wird in dieser Dissertation das Prinzip der linear-prädiktiven Codierung (LPC) verfolgt, wie es auch in der Sprachcodierung erfolgreich eingesetzt wird. Um möglichst alle oben genannten Randbedingungen erfüllen zu können, sind wesentliche Modifikationen erforderlich:

- Funktionale Komponenten zur Ausnutzung spezieller Eigenschaften von Sprachsignalen, wie zum Beispiel die Langzeitprädiktion (*Long Term Prediction*, LTP) zur Modellierung der momentanen Sprachgrundfrequenz bei der Sprachcodierung, werden nicht verwendet.

- Die bei der Sprachcodierung zur Quantisierung verwendeten, dünn besetzten Codebücher werden durch einen neuen Typ von Vektorquantisierer, den *Logarithmisch Sphärischen Vektorquantisierer* (LSVQ), ersetzt. Dies verbessert die erzielbare subjektive Audioqualität entscheidend.

- Ein neues Modell zur Stabilitäts- und Leistungsanalyse von *closed-loop* LPC (d.h. Quantisierung mit Fehlerrückführung) wird entwickelt, das nicht nur für hohe, sondern auch für niedrige Bitraten gültig ist.

- Ausgehend von dem neuen Modell wird ein neuartiges Optimierungskriterium zur Realisierung des von der *Rate Distortion Theorie* (RDT) bekannten *Reverse Waterfilling* abgeleitet, das bei der Codierung von Audiosignalen insbesondere bei niedrigen Bitraten und kurzen Blocklängen von großer Bedeutung ist.

- Anstelle der konventionellen linearen Prädiktion (LP) wird die sogenannte *Warped Linear Prediction* (WLP) eingesetzt. Die mit der WLP einhergehende ungleichmäßige Frequenzauflösung berücksichtigt die Eigenschaften der akustischen Wahrnehmung beim Menschen und führt somit – insbesondere bei Audiosignalen – zu einer verbesserten Klangqualität.

Die neu entwickelten Techniken werden zunächst theoretisch und teilweise sehr grundsätzlich betrachtet. Ein starker Praxisbezug ist dadurch gegeben, dass alle entwickelten Konzepte und Audiocodecs im Hinblick auf ihre praktische Einsetzbarkeit (insbesondere Rechen- und Speicherbedarf) untersucht und optimiert werden.

Grundlagen der Quantisierung

Im ersten Teil der vorliegenden Dissertation geht es um grundsätzliche Ausführungen zum Thema Quantisierung. Einführend werden zunächst die im Rahmen der *Rate Distortion Theorie* bekannten theoretischen Leistungsgrenzen für die Quantisierung unkorrelierter stationärer Quellen beschrieben. Daran anschließend werden die Ergebnisse auf korrelierte Quellen erweitert. Insbesondere der Einfluss des *Reverse Waterfilling* Prinzips wird genauer untersucht.

Im darauf folgenden Teil schließt sich eine Betrachtung der *asymptotischen Hochraten-Quantisierungstheorie* (Quantisierung mit hohen Bitraten) an. Es wird gezeigt, dass die Vektorquantisierung (VQ) der skalaren Quantisierung (SQ) durch die sogenannten *Vektorquantisierungs-Vorteile* überlegen ist. Eine konkrete Antwort auf die Frage nach einer praktischen Umsetzung eines optimalen Vektorquantisierers kann die Hochraten-Quantisierungstheorie nicht liefern. Die Bestimmungsgleichung der optimalen Codevektordichte aus der multivariaten Verteilungsdichtefunktion eines zu quantisierenden (Vektor-) Signals ist jedoch diesbezüglich ein wichtiger Hinweis.

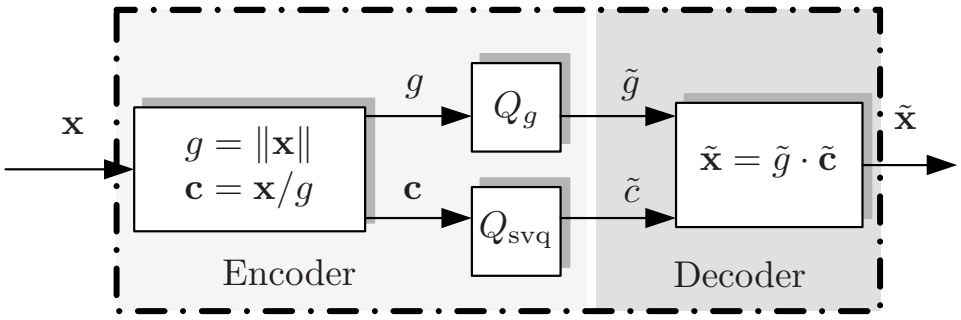


Abbildung C.1: Prinzip der Logarithmisch Sphärischen Vektorquantisierung.

Logarithmisch Sphärische Vektorquantisierung

Basierend auf den Ergebnissen der asymptotischen Hochraten-Quantisierungstheorie wird in der vorliegenden Arbeit gezeigt, dass eine Anordnung von Codevektoren auf der Oberfläche von (skalierten) mehrdimensionalen Sphären bei der *Logarithmisch Sphärischen Vektorquantisierung* (LSVQ) eine gute Approximation der optimalen Codevektordichte für den Fall einer multivariaten unabhängig und identisch verteilten Gauß'schen Quelle ist.

In der **Sprachcodierung** ist die Annahme einer Gauß'schen Verteilung bekanntermaßen eine gute Approximation für normierte Restsignale nach linearer Prädiktion. Bei der **Audiocodierung** dient diese Annahme hingegen in erster Linie der Vorgabe des ungünstigsten Szenarios (*worst case*), da die Gauß'sche Verteilung die höchste differentielle Entropie aufweist und deswegen in Bezug auf die Quantisierung die am „schwierigsten" zu behandelnde Verteilung darstellt. Praktisch führt dies dazu, dass eine minimale Audioqualität für beliebige zu quantisierenden Signale garantiert wird, was von Vorteil ist, da Audiosignale eine große Vielfalt unterschiedlicher Eigenschaften aufweisen können.

Das Prinzip der Logarithmisch Sphärischen Vektorquantisierung ist in Abbildung C.1 dargestellt. Dabei wird jeder zu quantisierende Vektor $\mathbf{x}$ in seinen Betrag g und einen normierten Vektor $\mathbf{c}$ mit dem Absolutbetrag $\|\mathbf{c}\| = 1$ zerlegt. Beide Komponenten werden anschließend mithilfe eines logarithmischen skalaren Quantisierers Q_g sowie eines sphärischen Vektorquantisierers Q_{svq}, bei dem alle Codevektoren möglichst gleichmäßig auf der Oberfläche einer Sphäre mit dem Radius 1 verteilt sind, quantisiert. Im Decoder kann der quantisierte Vektor $\tilde{\mathbf{x}}$ durch Multiplikation des quantisierten Betrages $\tilde{g}$ mit dem quantisierten normierten Vektor $\tilde{c}$ wieder rekonstruiert werden.

In der vorliegenden Dissertation werden neuartige *qualitative* theoretische Analysen des LSVQ Ansatzes für die Annahme hoher Bitraten abgeleitet. Hierfür wird ein analytischer Ausdruck zur Berechnung des Volumens der sich ergebenden Quantisierungszellen bestimmt, mit dem das Signal-zu-Rausch-Verhältnis bezüglich des rekonstruierten Eingangssignals (Quantisierungs-SNR) berechnet wird. Annahmen

über die Form der Quantisierungszellen werden nicht gemacht, so dass eine quantitative Auswertung der abgeleiteten Gleichungen aufgrund weniger unbekannter Konstanten nicht möglich ist. Als Schlussfolgerung aus diesen Analysen wird gezeigt, dass das Quantisierungs-SNR unabhängig von der Verteilung des zu quantisierenden Signals ist. Die Forderung nach einer minimal erreichbaren Qualität für alle Typen von Signalen ist somit auch mathematisch zu belegen.

Im nächsten Schritt wird ein *idealisierter* sphärischer Vektorquantisierer (SVQ) definiert. Annahmen über die Form der Quantisierungszellen ähnlich denen bei der von der Hochraten-Quantisierungstheorie bekannten *Sphere Upper Bound* ermöglichen die Bestimmung einer oberen Grenze für das durch LSVQ erreichbare Quantisierungs-SNR. Mithilfe der berechneten mathematischen Zusammenhänge und numerischer Verfahren kann diese Grenze für LSVQ für beliebige Vektordimensionen und Bitraten nun erstmals *qualitativ* berechnet werden. Die Einführung von Näherungen für die Annahme hoher Bitraten ermöglicht schließlich die Bestimmung der optimalen Allokation von Bits für den logarithmischen skalaren Quantisierer auf der einen Seite und den sphärischen Vektorquantisierer auf der anderen Seite, vorausgesetzt, dass ein festes Gesamtbitbudget pro Vektor für die Quantisierung vorgeben ist.

In einer abschließenden Auswertung werden unter anderem die folgenden Schlussfolgerungen gezogen:

- Das Quantisierungs-SNR des LSVQ Ansatzes steigt mit wachsender Vektordimension.
- Der Anstieg ist besonders stark für niedrige Vektordimensionen und wird mit zunehmender Dimension immer weniger ausgeprägt.
- LSVQ ist asymptotisch optimal für unendliche Vektordimensionen und stationäre Signale mit multivariater unabhängig und identischer Gauß'scher Verteilung.

Um über die Theorie hinaus einen Bezug zur Praxis herzustellen, werden drei Verfahren zur praktischen Umsetzung des LSVQ Konzepts und insbesondere des sphärischen Vektorquantisierers vorgestellt, der *Gosset Low Complexity Vector Quantizer* (GLCVQ), der *Algebraic Low Bitrate Vector Quantizer* (ALBVQ) und der *Apple Peeling Vector Quantizer* (APVQ). Neuartige Algorithmen zur *Nearest-Neighbor*-Quantisierung mit geringer Rechenkomplexität sowie zur kompakten Speicherung von Vektorcodebüchern werden entwickelt.

Schließlich wird die Leistungsfähigkeit aller drei Ansätze in Simulationen untersucht und mit der theoretischen Grenze für LSVQ verglichen. Es stellt sich heraus, dass alle drei Ansätze die theoretisch maximal möglich Leistungsgrenze annähernd erreichen und durch die aus der Strukturierung der Codebücher resultierende effiziente Umsetzbarkeit hervorragend für praktische Anwendungen geeignet sind.

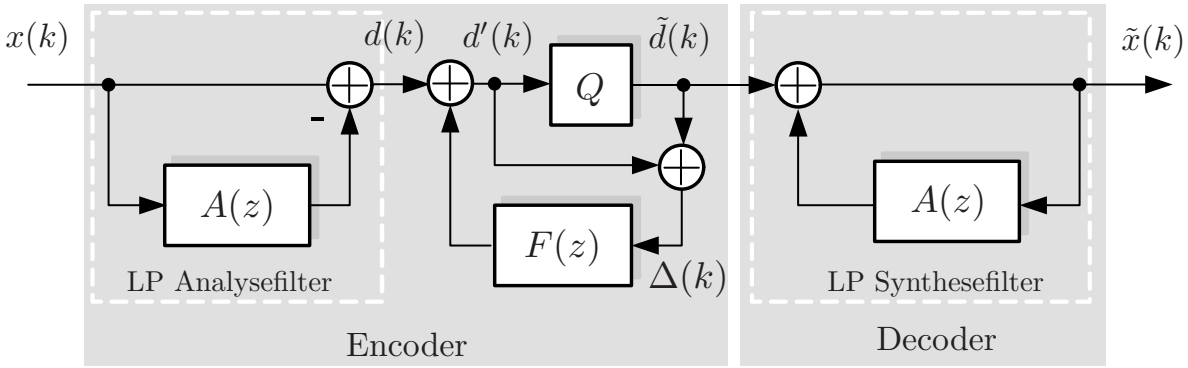


Abbildung C.2: Prinzip der kombinierten linearen Prädiktion und skalaren Quantisierung in *closed-loop* Struktur, Encoder und Decoder.

Linear-prädiktive Codierung

Der Ansatz der Logarithmisch Sphärischen Vektorquantisierung ist sehr leistungsfähig für unkorrelierte Quellen, kann aber die Korrelation eines zu quantisierenden Signals nicht zur Steigerung des Quantisierungs-SNRs nutzen. Aus diesem Grund wird der Quantisierer mit dem Prinzip der linearen Prädiktion (LP) kombiniert, was als *linear-prädiktive Codierung* (LPC) bezeichnet wird. In der Regel werden beide Teilkomponenten in einer geschlossenen Schleife (*closed-loop*) miteinander verbunden. Das Prinzip ist in allgemeiner Form in Abbildung C.2 dargestellt. Das *LP Analysefilter* im Encoder dient dazu, das Eingangssignal $x(k)$ vor der Quantisierung zu dekorrelieren. Die entsprechenden Filterkoeffizienten (LP Koeffizienten) werden in der Regel blockweise bestimmt und als Nebeninformation mit verhältnismäßig geringer Bitrate an den Decoder übertragen, wo sie zur Rekonstruktion des Eingangssignals durch das *LP Synthesefilter* benötigt werden.

In der allgemeinen *closed-loop* Struktur mit **skalarem Quantisierer** wird eine durch das Fehlergewichtungsfilter ($F(z)$) gefilterte Version des Quantisierungsrauschens rückgekoppelt (siehe Abbildung C.2).

Im Falle eines **Vektorquantisierers** ist die blockweise Verarbeitung der Vektoren einerseits und eine Filterung zur Fehlerrückführung Abtastwert für Abstastwert andererseits widersprüchlich. Aus diesem Grunde wird hierbei auf das *Code-Excited Linear Prediction* (CELP) Prinzip zurückgegriffen, bei dem die Quantisierung nach dem sogenannten *Analyse-durch-Synthese*-Prinzip realisiert wird.

Das Analyse-durch-Synthese-Prinzip ist typischerweise mit einem sehr hohen Rechenaufwand verbunden, da alle Einträge des umfangreichen Vektorcodebuchs zunächst gefiltert und anschließend mit dem unquantisierten Eingangssignal verglichen werden müssen, um den optimalen Codevektor zu bestimmen. In der vorliegenden Dissertation wird zunächst ein modifizierter CELP Ansatz abgeleitet, der die Einführung von Maßnahmen zur Komplexitätsreduktion bei der praktischen Umsetzung der beiden neu entwickelten Codecs, SCELP und W-SCELP, ermöglicht.

Darüber hinaus wird auf neuartige Weise nachgewiesen, dass der modifizierte CELP Ansatz für den Spezialfall einer Vektordimension von eins identisch mit der in Abbildung C.2 dargestellten Struktur ist.

Ein Ansatz zur theoretischen Beschreibung der Kombination von linearer Prädiktion und Quantisierung in (verallgemeinerter) *closed-loop* Struktur ist aus der Literatur bekannt, basiert jedoch auf der Annahme hoher Bitraten. Da in der Praxis eine Abweichung von der Theorie für niedrige Bitraten festgestellt wurde, wird in dieser Dissertation eine neue Theorie entwickelt, die auch bei niedrigen Bitraten Gültigkeit hat. Die Theorie basiert auf einem neuen skalaren Ansatz zur Modellierung der durch die Quantisierung erzeugten Signalverzerrungen, in dem insbesondere die Interaktion zwischen der Rückführung des Quantisierungsfehlers und dem Quantisierer selbst berücksichtigt wird.

Das neue Modell bestätigt die Ergebnisse der aus der Literatur bekannten Theorie für den Fall hoher Bitraten. Es zeigt jedoch auch, dass bei niedrigen Bitraten der Encoder instabil werden kann und mit einer deutlich geringeren Leistungsfähigkeit zu rechnen ist als bislang in der Literatur angegeben. Dieses Verhalten wurde zwar teilweise in der Literatur beschrieben, konnte jedoch erst durch das neue Modell auch theoretisch belegt werden.

Basierend auf den theoretischen Ergebnissen wird anschließend ein neuartiges Optimierungskriterium für die Berechnung der Koeffizienten des LP Analyse- und des Fehlergewichtungsfilters für niedrige Bitraten bestimmt. In einem aus zwei Schritten bestehenden Verfahren kann die Bestimmung der Koeffizienten unter Berücksichtigung des neuen Kriteriums praktisch sehr elegant und ohne großen zusätzlichen Rechenaufwand umgesetzt werden. Eine detaillierte Untersuchung des neuen Optimierungskriteriums zeigt, dass der Zwei-Schritt-Ansatz optimal in dem Sinne ist, dass das *Reverse Waterfilling* Prinzip, das in der Rate Distortion Theorie bei der Betrachtung korrelierter Signale von großer Bedeutung ist, hier nun erstmalig bei der LPC Anwendung findet.

Die aufgezeigten Zusammenhänge sind bei der Codierung von Audiosignalen und im Falle kurzer LP Analyse-Blocklängen von entscheidender Bedeutung. Aus diesem Grund haben sie einen wesentlichen Einfluss auf den Entwurf von Verfahren zur Audiocodierung mit geringer Verzögerung. Neben dieser praktischen Relevanz erlaubt die im Rahmen dieser Arbeit formulierte Theorie neue Einblicke in das grundlegende Verständnis der linear-prädiktiven Codierung.

Der SCELP und der W-SCELP Audiocodec

Im letzten Teil der Dissertation werden die zuvor entwickelten Techniken praktisch angewendet: Die Logarithmisch Sphärische Vektorquantisierung wird mit der linearen Prädiktion auf Basis des modifizierten CELP Ansatzes kombiniert und bildet

so die Basis für den SCELP Audiocodec mit geringer algorithmischer Verzögerung. Zum Einsatz kommen kann prinzipiell jeder der drei entwickelten LSVQ Ansätze. Da, wie in der Arbeit gezeigt wird, die optimierten Verfahren zur *Nearest-Neighbor*-Quantisierung durch die Kombination mit der linearen Prädiktion nicht verwendet werden können, spielt für den praktischen Einsatz das Erreichen einer minimalen Rechenkomplexität des Analyse-durch-Synthese-Prinzips im CELP Encoder eine besondere Rolle.

In diesem Sinne ist von den drei entwickelten Ansätzen für LSVQ der APVQ besonders geeignet. Der Grund für diesen Vorzug liegt darin begründet, dass durch Ausnutzung der Eigenschaften der Codevektoren des APVQ Codebuchs eine sehr effiziente Variante der Analyse-durch-Synthese basierten Codevektorsuche (d.h. der Quantisierung) umgesetzt werden kann. Basis für diese effiziente Umsetzung ist die Entwicklung von Verfahren zur Komplexitätsreduktion, die in der vorliegenden Dissertation vorgestellt werden. Im Vergleich zu einer Analyse-durch-Synthese Codevektorsuche, bei der alle Einträge des Codebuchs getestet werden, um den optimalen Codevektor zu bestimmen, erlauben es diese Maßnahmen zur Komplexitätsreduktion, die theoretische Rechenkomplexität um einen Faktor von bis zu 10^5 zu reduzieren. Die Audioqualität wird dabei je nach Komplexitätseinsparung nicht wahrnehmbar oder nur geringfügig verschlechtert.

Um darüber hinaus eine höhere subjektive Qualität insbesondere für Audiosignale zu erreichen, wird der SCELP Audiocodec schließlich durch die Verwendung der *Warped Linear Prediction* (WLP) anstelle der konventionellen linearen Prädiktion verbessert. Grundlage dieser Verbesserung ist eine ungleichmäßige Frequenzauflösung mittels WLP, welche einer gleichförmigen Auflösung überlegen ist, da die Eigenschaften der menschlichen akustischen Wahrnehmung besser abgebildet werden.

Der SCELP und der W-SCELP Audiocodec verwenden nicht die Eigenschaften eines speziellen Signalmodells zur Erhöhung der Codiereffizienz. Aus diesem Grund ist das Gesamtkonzept hochgradig skalierbar und vielseitig einsetzbar. Die Umsetzung einer *Standard-Konfiguration* in Festkomma-Arithmetik mit Bestimmung der benötigten Rechenkomplexität[1] sowie eine Realisierung der Codecs als Echtzeit Prototyp auf einem PC belegen, dass die maximalen Rechenanforderungen moderat sind, so dass das vorgeschlagene Konzept auch in der Praxis gut einsetzbar ist.

Um die erreichbare Sprach- und Audioqualität zu bewerten, werden die objektiven Qualitätsmaße WB-PESQ (*Wideband Perceptual Speech Quality*) für Sprach- und PEAQ (*Perceptual Audio Quality*) für Audiosignale mit einer Audiobandbreite von 50-7000 Hz (Breitbandsprache/Audio, Abtastrate $f_s = 16$ kHz) verwendet. Für eine Beurteilung und qualitative Einordnung der vorgestellten Codecvarianten

[1] Encoder: 28 WMOPS, Decoder: 3 WMOPS (WMOPS: Weighted Million Operations per Second)

SCELP und W-SCELP werden zwei grundsätzlich unterschiedliche Vergleichscodecs für Breitbandsprache/Audio betrachtet:

- Der ITU-T G.722 Codec, der bei den Bitraten 48, 56 und 64 kBit/sec betrieben wird und eine algorithmische Verzögerung von weniger als 10 ms aufweist.
- Der MPEG I, *audio layer* 3, (MP3) Codec, der bei einer Abtastrate von 16 kHz mit den Bitraten 24 und 32 kBit/sec betrieben wird und eine algorithmische Verzögerung von mehr als 80 ms aufweist.

Der Bewertungsvergleich für Audiosignale ist in Abbildung C.3 zusammengefasst. Die Auswertung der Qualitätsbewertung der Codecs mit den objektiven Maßen

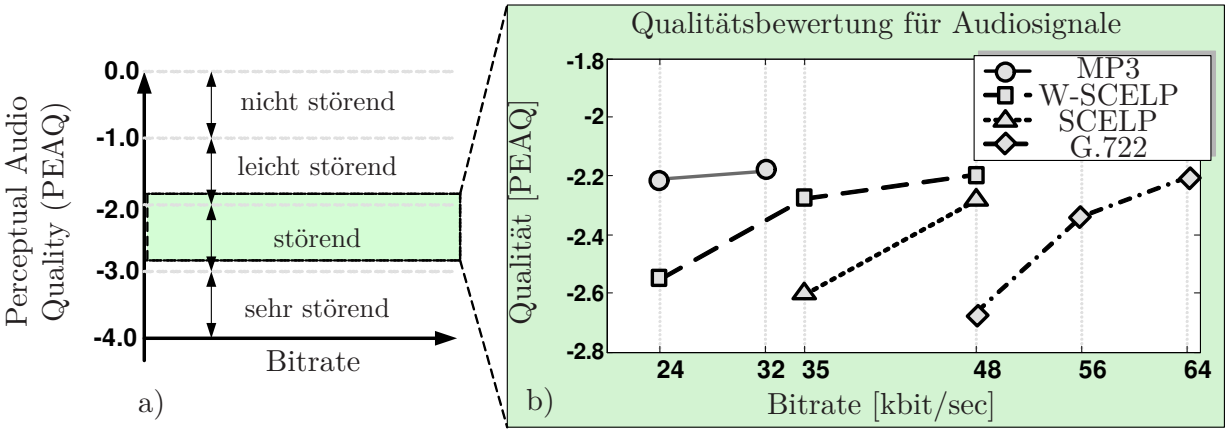


Abbildung C.3: Klassifizierung der Qualität von Audiosignalen nach dem PEAQ (a) sowie gemessene PEAQ Kennzahlen für den W-SCELP, den SCELP, den ITU-T G.722 und den MP3 Audiocodec (b). Verzögerung ITU-T G.722, W-SCELP, SCELP Codec < 10 ms, Verzögerung MP3 Codec > 80 ms.

zeigt, dass der SCELP und der W-SCELP Audiocodec dem ITU-T G.722 Codec in Bezug auf die erreichbare Qualität insbesondere für Audiosignale deutlich überlegen sind. Gegenüber dem MP3 Codec bei 32 kBit/sec erzielt der W-SCELP Codec für eine Bitrate von 35 kBit/sec annähernd vergleichbare Audioqualität. Bei der Beurteilung ist jedoch zu berücksichtigen, dass der W-SCELP Codec mit weniger als 10 ms eine um den Faktor 8 geringere algorithmische Verzögerung aufweist.

Der Vergleich aller Ergebnisse zeigt, dass für Anwendungen wie Drahtlos-Mikrofone bei Live-Konzerten oder drahtlose Audioverbindungen in Hörgeräten etc. nur der SCELP und der W-SCELP Codec ausreichende Sprach- und insbesondere Audioqualität bei gleichzeitiger Einhaltung der eingangs geforderten engen technischen Randbedingungen liefern können.

Bibliography

[AB88] J. Adoul and M. Barth. "Nearest Neighbor Algorithm for Spherical Codes from the Leech Lattice". *IEEE Transactions on Information Theory*, 34(5):1188–1202, 1988.

[Abu90] H. Abut. *Vector Quantization*. IEEE Reprint Collection, IEEE Press, New Jersey, May 1990.

[Ado07] Adobe System Incorporated. "Adobe Audition". http://www.adobe.com, 2007.

[AL87] J.-P. Adoul and C. Lamblin. "A Comparison of some Algebraic Structures for CELP Coding of Speech". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1987.

[ALL84] J. P. Adoul, C. Lamblin, and A. Leguyader. "Baseband Speech Coding at 2400 bps using Spherical Vector Quantization". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, March 1984.

[Ari72] S. Arimoto. "An Algorithm for calculating the capacity of an arbitrary discrete memoryless channel". *IEEE Transactions on Information Theory*, 1972.

[AS67] B. Atal and M. Schroeder. "Predictive Coding of Speech Signals". *Proc. Conf. Commun. and Process.*, pages 360–361, 1967.

[AS84] B. S. Atal and M. R. Schroeder. "Stochastic coding of speech signals at very low bit rates". *Proc. Int. Conf. Communications-ICC*, pages 1610–1613, 1984.

[Ast84] J. Astola. "The Tietavainen Bound for Spherical Codes". *Discrete Appl. Math.*, 7(1):17–21, 1984.

[Ata82] B. Atal. "Predictive Coding of Speech at Low bit Rates". *IEEE Transactions on Communications*, COM-30(4):600–614, apr 1982.

[Ata86] B. S. Atal. "High-quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 1681–1684, 1986.

[Bad06] A. Badach. *Voice over IP - Die Technik: Grundlagen, Protokolle, Anwendungen, Migration, Sicherheit (Gebundene Ausgabe)*. Hanser Fachbuchverlag, 2006.

[Ben48] W. R. Bennett. "Spectra of Quantized Signals". *Bell Systems Techn. Journal*, 1948.

[Ber71] T. Berger. *Rate Distortion Theory*. Prentice-Hall Electrical Engineering Series, 1971.

[Ber72] T. Berger. "Optimum Quantizers and Permutation Codes". *IEEE Transactions on Information Theory*, IT-17:759–165, November 1972.

[BGK+08] S. Bruhn, V. Grancharov, W. B. Kleijn, J. Klejsa, M. Li, J. Plasberg, H. Pobloth, S. Ragot, and A. Vasilache. "The FlexCode Speech and Audio Coding Approach". *ITG Fachtagung Sprachkommunikation*, October 2008.

[BJ79] J. A. Bucklew and N. C. G. Jr. "Quantization schemes for bivariate Gaussian random variables". *IEEE Transactions on Information Theory*, IT-25:537–543, September 1979.

[BJW74] H. Brehm, E. W. Jüngst, and D. Wolf. "Simulation von Sprachsignalen". *Archiv für Elektronik und Übertragungstechnik (AEÜ)*, (28):445–450, 1974.

[Bla72] R. Blahut. " Computation of channel capacity and rate-distortion functions". *IEEE Transactions on Information Theory*, 80(4):460–473, 1972.

[Bla04] K. Blanchette. *Effects of MP3 Technology on the Music Industry: An Examination of Market Structure and Apple iTunes*. PhD thesis, Holy Cross, 2004.

[Bli35] H. Blichfeldt. "The minimum values of positive quadratic forms in six, seven and eight variables". *Mathematische Zeitschrift*, 39:1–15, 1935.

[Bra06] K. Brandenburg. "Perceptual Coding of High Quality Digital Audio". *Applications of Digital Signal Processing to Audio and Acoustics*, pages 39–83, 2006.

[BS91] I. N. Bronstein and K. A. Semendjajew. *Taschenbuch der Mathematik*. Teubner Verlag, 1991.

[Cat69] K. W. Cattermole. *Principles of Pulse Code Modulation*. Iliffe Books London, 1969.

[Coo52] J. Coolidge. "The Origin of Polar Coordinates". *American Mathematical Monthly*, (59):78–85, 1952.

[Cox68] H. Coxeter. *Twelve Geometric Essays*. Southern Illinois University Press, 1968.

[Cox73] H. S. M. Coxeter. *Regular Polytopes, 3rd ed.* New York: Dover, 1973.

[CR83] R. E. Crochiere and L. R. Rabiner. *Multirate Digital Signal Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1983.

[CS85] J. H. Conway and N. J. A. Sloane. "A Lower Bound on the Average Error of Vector Quantizers". *IEEE Transactions on Information Theory*, IT-31(1):106–109, 1985.

[CS93] J. H. Conway and N. J. A. Sloane. *Sphere Packings, Lattices and Groups*. Springer-Verlag, New York, 1993.

[Csi74] I. Csiszar. "On the Computation of Rate Distortion Functions". *IEEE Transactions on Information Theory*, 1974.

[CT91] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley Verlag, Berlin, 1991.

[Dur60] J. Durbin. "The Fitting of Time-Series Models". *Revue de l'Institute International de Statistique*, 28(3):233–243, 1960.

[dWB00] S. de Waele and P. M. T. Broersen. "The Burg Algorithm for Segments". *Proc. of Intl. Conf. on Signal Processing (ICSP)*, 48(10):2876–2880, October 2000.

[Erd04] C. Erdmann. *Hierarchical Vector Quantization: Theory and Application to Speech Coding*. PhD thesis, RWTH Aachen, 2004.

[ETS96] ETSI, Rec. GSM 06.60. "Enhanced Full Rate Speech Transcoding", 1996.

[ETS00] ETSI, Rec. GSM 06.90. "Digital Cellular Telecommunications System (Phase 2+); Adaptive Multi-Rate (AMR) Speech Transcoding". version 7.2.1, release 1998, April 2000.

[ETS01] ETSI/3GPP, Rec. GSM 26.190/ITU-T G.722.2. "Adaptive Multi-Rate Wideband Speech Transcoding (AMR-WB)", 2001.

[ETS05] ETSI, Rec. GSM 26.290. "Extended Adaptive Multi-Rate - Wideband (AMR-WB+) Codec", 2005.

[Fan60] G. Fant. *Acoustic Theory of Speech Production*. Mouton and Co.'s-Gravenhage, The Netherlands, 1960.

[Fan61] R. M. Fano. *Transmission of Information: A Statistical Theory of Communicaton*. Wiley, New York, 1961.

[FL84] N. Farvardin and F. Y. Lin. "Optimal Quantizers for a class of non-Gaussian memoryless sources". *IEEE Transactions on Information Theory*, IT-30:485–497, may 1984.

[Fla72] J. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[FR00] N. H. Fletcher and T. D. Rossing. *The Physics of Musical Instruments*. Springer, Berlin, 2000.

[Fra07] Fraunhofer Institut Digitale Medientechnologie. "Audio Coding with Ultra Low Encoding/Decoding Delay". http://www.idmt.fraunhofer.de, 2007.

[Ger72] A. Gersho. "Stochastic stability of delta modulation". *Bell Syst. Tech. Jour.*, 51:821–841, April 1972.

[Ger79] A. Gersho. "Asymptotically Optimal Block Quantization". *IEEE Transactions on Information Theory*, 25(4):373–380, 1979.

[GH67] T. J. Goblick and J. Holsinger. "Analog Source Digitization: A Comparison of Theory and Practice". *IEEE Transactions on Information Theory*, IT-13:323–326, apr 1967.

[GHSW87] E. Gamal, L. Hemachandra, I. Spherling, and V. Wei. "Using Simulated Annealing to Design Good Codes". *IEEE Transactions on Information Theory*, 33(1), 1987.

[GKL+09] B. Geiser, H. Krüger, H. W. Löllmann, P. Vary, D. Zhang, H. Wan, H. T. Li, and L. B. Zhang. "Candidate Proposal for ITU-T Super-wideband Speech and Audio Coding". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4121–4124, Taipei, Taiwan, 2009.

[GL77] L. H. Goldstein and B. Liu. "Deterministic and Stochastic Stability of Adaptive Differential Pulse Code Modulation". *IEEE Transactions on Information Theory*, 23(4):445–453, July 1977.

[GN98] R. M. Gray and D. Neuhoff. "Quantization". *IEEE Transactions on Information Theory*, 1998.

[Gos00] T. Gosset. "On the regular and semi-regular figures in space of n dimensions". *Messenger of Mathematics*, 29:43–48, 1900.

[GP68] H. Gish and J. N. Pierce. "Asymptotically Efficient Quantization". *IEEE Transactions on Information Theory*, IT-14:676–683, September 1968.

[GR92] A. Gersho and R.M.Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

[GS58] U. Grenander and G. Szego. *Toeplitz Forms and their Applications*. Univ. of California Press, 1958.

[GS88] J. D. Gibson and K. Sayood. "Lattice Quantization". *Advances in Electronics and Electron Physics*, 72:259–330, 1988.

[HÖ98] A. Härmä. "Implementation of Recursive Filters Having Delay Free Loops". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1261–1264, Seattle, USA, May 1998.

[HÖ00] A. Härmä. "Implementation of Frequency-Warped Recursive Filters". *Signal Processing, Elsevier*, 80(3):1986–1990, March 2000.

[Ham96] J. Hamkins. *Design and Analysis of Spherical Codes*. PhD thesis, University of Illinois, 1996.

[Hei01] S. Heinen. *Quellenoptimierter Fehlerschutz für Digitale Übertragungskanäle*. PhD thesis, RWTH Aachen, 2001.

[HL99] A. Härmä and U. Laine. "Warped low-delay CELP for Wide-Band Audio Coding". *Proc. AES 17th Int. Conf. High-Quality Audio Coding*, September 1999.

[HL01] A. Härmä and U. Laine. "A Comparison of Warped and Conventional Linear Predictive Coding". *IEEE Trans. Speech and Audio Processing*, 9(5), 2001.

[HM04] J. B. Huber and B. Matschkal. "Spherical logarithmic Quantization and its Application for DPCM". *5th International ITG Conference on Source and Channel Coding (SCC)*, pages 349–356, Erlangen, Germany, jan 2004.

[Hol49] H. Holzwarth. "Pulse Code Modulation und ihre Verzerrungen bei logarithmischer Amplitudenquantelung". *Archiv der elektrischen Übertragung*, 1949.

[HS63] J. J. Y. Huang and P. M. Schultheiss. "Block Quantization of Correlated Gaussian - Random Variables". *IEEE Trans. on Communication Systems*, 11(9):289–296, September 1963.

[HSW01] L. Hanzo, F. Sommerville, and J. Woodard. *Voice Compression and Communications*. Wiley-Interscience, New York, 2001.

[Huf52] D. Huffman. "A method for the construction of minimum redundancy codes". *Proc. IRE*, 40:1098–1101, sep 1952.

[HZ97a] J. Hamkins and K. Zeger. "Asymptotically Dense Spherical Codes - Part I: Wrapped Spherical Codes". *IEEE Transactions on Information Theory*, 43(6):1774–1785, 1997.

[HZ97b] J. Hamkins and K. Zeger. "Asymptotically Dense Spherical Codes - Part II: Laminated Spherical Codes". *IEEE Transactions on Information Theory*, 43(6):1786–1798, 1997.

[HZ03] J. Hamkins and K. Zeger. "Gaussian Source Coding with Spherical Codes". *IEEE Transactions on Information Theory*, 48(11):2980–2989, 2003.

[ISO93] ISO/IEC 11172-3. "Information Technology – Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s – Part 3: Audio ", 1993.

[ISO97] ISO/IEC 13818-7. "Information Technology – Generic Coding of Moving Pictures and Associated Audio Information – Part 7: Advanced Audio Coding (AAC)", 1997.

[ISO05] ISO/IEC 14496-3. "Information Technology – Coding of Audio-Visual Objects – Part 3: Audio", 2005.

[Ita75] F. Itakuro. "Line Spectral Representation of Linear Prediction Coefficients of Speech Signals". *Journal of the Acoustical Society of America*, 57(1):35, 1975.

[ITU88a] ITU-T, Rec. G.711. "Pulse Code Modulation (PCM) of Voice Frequencies", 1988.

[ITU88b] ITU-T, Rec. G.722. "7 kHz Audio Coding within 64 kbit/s", 1988.

[ITU88c] ITU-T, Rec. I.241-1. "Teleservices supported by an ISDN: Telephony", November 1988.

[ITU92] ITU-T, Rec. G.728. "Coding of speech at 16 kbit/s using low-delay code excited linear prediction", 1992.

[ITU93] ITU-T, Rec. I.120. "Integrated services digital networks (ISDNs)", March 1993.

[ITU96] ITU-T, Rec. G.729. "Coding of Speech at 8 kbit/s using Conjugate-structure Algebraic-code-excited Linear Prediction (CS-ACELP)", March 1996.

[ITU98] ITU-R Rec. BS.1387. "Method for Objective Measurements of Perceived Audio Quality", December 1998.

[ITU00] ITU-T Rec. P.191. "Software Tools for Speech and Audio Coding Standardization", 2000.

[ITU05] ITU-T Rec. P.862.2. "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs", 2005.

[ITU06] ITU-T, Rec. G.729.1. "G.729 based Embedded Variable Bit-Rate Coder: An 8-32 kbit/s Scalable Wideband Coder Bitstream interoperable with G.729 ", 2006.

[IX89] M. Ireton and C. Xydeas. "On improving vector excitation coders through the use of spherical lattice codebooks (SLCs)". *ICASSP*, pages 57–60, 1989.

[Jel68] F. Jelinek. "Buffer Overflow in Variable Length Coding of Fixed Rate Sources". *IEEE Transactions on Information Theory*, IT-14(3):490–501, may 1968.

[JN84] N. Jayant and P. Noll. *Digital Coding of Waveforms*. Prentice-Hall, Inc., 1984.

[Kab02] P. Kabal. "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality". *TSP Lab Technical Report, Dept. Electrical and Computer Engineering, McGill University, http://www.tsp.ece.McGill.ca/MMSP/Documents*, may 2002.

[Kar47] K. Karhunen. "Über lineare Methoden in der Wahrscheinlichkeitsrechnung". *Ann. Academic Sciences Fennicae. Ser. A. I. Math.-Phys.*, (37):1–79, 1947.

[KD83] J. C. Kieffer and J. G. Dunham. "On a Type of Stochastic Stability Class of Encoding Schemes". *IEEE Transactions on Information Theory*, 29(6):793–797, November 1983.

[Kei06] F. Keiler. *Beiträge zur Audiocodierung mit kurzer Latenzzeit*. PhD thesis, Helmut-Schmidt-Universität/Universität der Bundeswehr Hamburg, 2006.

[KGV08] H. Krüger, B. Geiser, and P. Vary. "Method and Apparatus of Communication - Patent PCT/CN2008/070719 (pending)", 2008.

[Kie82] J. C. Kieffer. "Stochastic Stability for Feedback Quantization Schemes". *IEEE Transactions on Information Theory*, 28(2):248–254, March 1982.

[KJLV09] H. Krüger, M. Jeub, H. W. Löllmann, and P. Vary. "RTPROC: Rapid Real-Time Prototyping for Audio Signal Processing". *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Taipei, Taiwan, April 2009. IEEE. Show and Tell Demonstration.

[KKO08] J. Kim, M. Kim, and E. Oh. "Framework for Unified Speech and Audio Coding". *34th International Conference: New Trends in Audio for Mobile and Handheld Devices*, August 2008.

[KLEV03] H. Krüger, T. Lotter, G. Enzner, and P. Vary. "A PC based Platform for Multichannel Real-time Audio Processing". *Proc. of Intl. Workshop on Acoustic Echo and Noise Control (IWAENC)*, Kyoto, Japan, September 2003.

[KO07] W. B. Kleijn and A. Ozerov. "Rate Distribution between Model and Signal". *Proc. of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 360–361, 2007.

[Kos63] V. N. Koshelev. "Quantization with Minimum Entropy". *Probl. Pered. Inform.*, (14):151–156, 1963.

[KP95] W. B. Kleijn and K. K. Paliwal. *Speech Coding and Synthesis*. Elsevier, Amsterdam, 1995.

[Krü09] H. Krüger. "Low Delay Audio Coding based on Logarithmic Spherical Vector Quantization - Supplement to PHD Thesis", September 2009.

[KSE+09] H. Krüger, T. Schumacher, T. Esch, B. Geiser, and P. Vary. "RT-PROC: Rapid Real-Time Prototyping for Audio Signal Processing". *20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, Dresden, Germany, September 2009. TUDpress Verlag der Wissenschaften.

[KSGV08] H. Krüger, R. Schreiber, B. Geiser, and P. Vary. "On Logarithmic Spherical Vector Quantization". *Proceedings of International Symposium on Information Theory and its Applications (ISITA)*, Auckland, New Zealand, December 2008. Society of Information Theory and its Applications (SITA).

[KSV06] H. Krüger, L. Schmalen, and P. Vary. "Audio-Link for Hearing Aids - Project Report", October 2006.

[KSW+04] U. Krämer, G. Schuller, S. Wabnik, J. Klier, and J. Hirschfeld. "Ultra Low Delay audio coding with constant bit rate". *117th AES Convention*, October 2004.

[KV02] H. Krüger and P. Vary. "Entwicklung des JMEDIA Streaming Codecs - Projekt-Abschlussbericht", 2002.

[KV05] H. Krüger and P. Vary. "A Partial Decorrelation Scheme for Improved Predictive Open Loop Quantization with Noise Shaping". *Proc. of European Conf. on Speech Communication and Technology (EUROSPEECH)*, Lissabon, Portugal, September 2005.

[KV06a] H. Krüger and P. Vary. "An Efficient Codebook for the SCELP Low Delay Audio Codec". *Proc. of 8th Workshop on Multimedia Signal Processing (MMSP)*, Victoria, Canada, October 2006.

[KV06b] H. Krüger and P. Vary. "SCELP: Low Delay Audio Coding with Noise Shaping based on Spherical Vector Quantization". *Proc. of European Signal Processing Conf. (EUSIPCO)*, Florence, Italy, September 2006.

[KV07a] H. Krüger and P. Vary. "Warped Linear Prediction for Improved Perceptual Quality in the SCELP Low Delay Audio Codec (W-SCELP)". *Digital Audio Effects Conference (DAFX)*, Bordeaux, France, September 2007.

[KV07b] H. Krüger and P. Vary. "Method and Device for coding audio data based on vector quantisation - Patent EP1879179A1", 2007.

[KV08a] H. Krüger and P. Vary. "A New Approach for Low-Delay Joint-Stereo Coding". *ITG-Fachtagung Sprachkommunikation*, Aachen, Germany, October 2008.

[KV08b] H. Krüger and P. Vary. "A new Approach for Low-Delay Joint-Stereo Coding - Patent EP 08 012 311 and US 12/499,250 (pending)", 2008.

[KV08c] H. Krüger and P. Vary. "RTPROC: A System for Rapid Real-Time Prototyping in Audio Signal Processing". *Proceedings of IEEE/ACM International Symposium on Distributed Simulation and Real Time Applications*, Vancouver, BC, Canada, October 2008.

[LAMM89] C. Lamblin, J. P. Adoul, D. Massaloux, and S. Morissette. "Fast CELP Coding based on the Barnes-Wall Lattice in 16 Dimensions". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, May 1989.

[LAS+91] C. Laflamme, J. Adoul, R. Salami, S. Morisette, and P. Mabilleau. "16 kbps Wideband Speech Coding Technique Based on Algebraic CELP". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 13–16, 1991.

[LASM90] C. Laflamme, J.-P. Adoul, H. Y. Su, and S. Morissette. "On reducing computational complexity of codebook search in CELP coder through the use of algebraic codes". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 177–180, 1990.

[LBG80] Y. Linde, A. Buzo, and R. M. Gray. "An Algorithm for Vector Quantization Design". *IEEE Transactions on Communications*, 28(1):84–95, 1980.

[Lee64] J. Leech. "Some Sphere Packings in Higher Space". *Canad. J. Math.*, 16:657–682, 1964.

[Lee67] J. Leech. "Notes on Sphere Packings". *Canad. J. Math.*, 19:251–267, 1967.

[Lev47] N. Levinson. "The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction". *Journal of Mathematical Physics*, 25(4):261–278, 1947.

[LG89] T. D. Lookabough and R. Gray. “High-Resolution Quantization Theory and the Vector Quantizer Advantage”. *IEEE Transactions on Information Theory*, 35(5):1020–1083, 1989.

[Lük95] H. D. Lüke. *Signalübertragung. Grundlagen der digitalen und analogen Nachrichtenübertragungssysteme*. Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 6th edition, 1995.

[Llo82] S. Lloyd. “Least Squares Quantization in PCM”. *IEEE Transactions on Information Theory*, 28, 1982.

[LZ94] T. Linder and K. Zeger. “Asymptotic Entropy-Constrained Performance of Tessellating and Universal Randomized Lattice Quantization”. *IEEE Transactions on Information Theory*, 40(2):575–579, mar 1994.

[Mac03] D. J. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[Mak75] J. Makhoul. “Linear Prediction - A Tutorial Review”. *IEEE Proceedings*, 63:561–580, April 1975.

[Mat07] B. Matschkal. *Spherical Logarithmic Quantization*. PhD thesis, Erlanger Berichte aus Informations- und Kommunikationstechnik, 2007.

[Max60] J. Max. “Quantizing for Minimum Distortion”. *IRE Trans. Inform. Theory*, IT-6:7–12, mar 1960.

[MBH06] B. Matschkal, F. Bergner, and J. B. Huber. “Joint Signal Processing for Spherical Logarithmic Quantization and DPCM”. *Proc. of 4th International Symposium on Turbo Codes in connection with the 6th International ITG-Conference on Source and Channel Coding*, München, Germany, apr 2006.

[MG76] J. D. Markel and A. H. Gray jr. *Linear Prediction of Speech*. Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[MM84] K. Motoishi and T. Misumi. “On a Fast Vector Quantization Algorithm”. *Proc. VIIth Symp. Inform. Theory and its Applications*, 1984.

[Moo65] R. A. Moog. “A Voltage-Controlled Low-Pass High-Pass Filter for Audio Signal processing”. *Audio Engineering Society Copnvention*, 1965.

[MSG85] J. Makhoul, R. S, and H. Gish. “Vector Quantization in Speech Coding”. *Proc. of the IEEE*, 71(11):1551–1588, 1985.

[Neu96] D. L. Neuhoff. *Quantization Analysis (Theory of Lossy Source Coding)*. ICASSP’96 Tutorial, 1996.

[Nol74] P. Noll. "Adaptive Quantizing in Speech Coding Systems". *Proc. Intl. Zurich Seminar on Digital Communications*, 1974.

[NZ78] P. Noll and R. Zelinski. "Bounds on Quantizer Performance in the Low Bit-Rate Region". *IEEE Transactions on Communications*, COM-26:300–305, 1978.

[OS92] A. V. Oppenheim and R. W. Schafer. *Zeitdiskrete Signalverarbeitung*. R. Oldenbourg Verlag, München, Wien, 1992.

[PA93] K. Paliwal and B. Atal. "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1:3–13, 1993.

[Pau96] J. Paulus. *Codierung breitbandiger Sprachsignale bei niedriger Datenrate*. PhD thesis, RWTH Aachen, 1996.

[PB86] J. Princen and A. Bradley. "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation". *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(5):1153–1161, October 1986.

[PD51] P. Panter and W. Dite. "Quantizing Distortion in pulse-count modulation with non-uniform spacing of levels.". *Proc. IRE*, 1951.

[Pea79] W. A. Pearlman. "Polar quantization of a complex Gaussian random variable". *IEEE Transactions on Communications*, COM-27:892–899, June 1979.

[PS00] T. Painter and A. Spanias. "Perceptual Coding of Digital Audio". *Proceedings of the IEEE*, volume 88, no.4, pages 451–513. IEEE, 2000.

[Pud08] H. Puder. "Compensation of Hearing Impairment with Hearing Aids: Current Solutions and Trends". *ITG Fachtagung Sprachkommunikation*, October 2008.

[Ric73] D. Richards. *Telecommunication by Speech*. Butterworths, London, 1973.

[Ris76] J. Rissanen. "Generalized Kraft Inequality and Arithmetic Coding". *IBM Journal of Research and Development*, 20(2):198–203, may 1976.

[RS78] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice-Hall Signal Processing Series, Engelwood Cliffs, New Jersey, 1978.

[SA85] M. Schroeder and B. Atal. "Code-excited Linear Prediction (CELP): High-quality Speech at very Low Bit Rates". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1985.

[SA89] R. Salami and D. Appleby. “A new Approach to low Bit Rate Speech Coding with low Complexity using Binary Pulse Excitation (BPE)”. *IEEE Workshop on Speech Coding for Telec.*, September 1989.

[SA99] J. O. Smith III and J. S. Abel. “Bark and ERB Bilinear Transforms”. *IEEE Transactions on Speech and Audio Processing*, 7(6):697–708, November 1999.

[SAH79] M. Schroeder, B. Atal, and J. Hall. “Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear”. *Journal of the Acoustical Society of America*, pages 1647–1652, 1979.

[Sal89] R. Salami. “Binary code excited linear prediction (BCELP): New approach to CELP coding of speech without codebooks”. *IEE Electronic Letters*, 25:401–403, mar 1989.

[Sch06] R. Schreiber. “Efficient Low Delay Audio Coding”. Diplomarbeit, RWTH Aachen, Institut für Nachrichtengeräte und Datenverarbeitung, 2006.

[SG84] M. Sabin and R. Gray. “Product Code Vector Quantizers for Waveform and Voice Coding”. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32(3):474–488, jun 1984.

[SG93] D. Sampson and M. Ghanbari. “Fast Lattice-Based Gain-Shape Vector Quantization for Image-Sequence Coding”. *IEE Proceedings-I*, 140(1), 1993.

[Sha48] C. Shannon. “A Mathematical Theory of Communication”. *The Bell System Technical Journal*, 27:379–423, 1948.

[SKV09] M. Schäfer, H. Krüger, and P. Vary. “Extending Monaural Speech and Audio Codecs by Inter-Channel Linear Prediction”. *20. Konferenz Elektronische Sprachsignalverarbeitung (ESSV)*, volume 1, page 8, Dresden, Germany, September 2009. ITG, DEGA, TUDpress Verlag der Wissenschaften.

[SLAM94] R. Salami, C. Laflamme, J. Adoul, and D. Massaloux. “A Toll Quality 8kb/s Speech Codec for the Personal Communication System (PCS)”. *IEEE Trans. Vehicular Technology*, 43:808–816, 1994.

[Smi57] B. Smith. “Instantaneous Companding of Quantized Signals”. *Bell Systems Techn. Journal*, 1957.

[SS87] M. Schroeder and N. Sloane. “New Permutation Codes Using Hadamard Unscrambling”. *IEEE Transactions on Information Theory*, IT-33(1):144–146, jan 1987.

[ST83] P. F. Swaszek and J. B. Thomas. "Multidimensional Spherical Coordinates Quantization". *IEEE Transactions on Information Theory*, IT-29(4):570–577, July 1983.

[Str80] H. W. Strube. "Linear Prediction on a Warped Frequency Scale". *Journal of the Acoustical Society of America*, 68:1071–1076, 1980.

[T.C06] T.Clevorn. *Turbo DeCodulation: Iterative Joint Source-Channel Decoding and Demodulation*. Number 24 in Aachener Beiträge zu Digitalen Nachrichtensystemen. nov 2006.

[Thi00] T. Thiede. "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality". *J. Audio Eng. Soc.*, (48):3–29, jan 2000.

[Tot59a] L. F. Toth. "Kugelunterdeckungen und Kugelüberdeckungen in Räumen konstanter Krümmung". *Archiv Math.*, 10:307–313, 1959.

[Tot59b] L. F. Toth. "Sur la representation d'une population infinie par un nombre fini d'elements". *Acta Math. Acad. Scient. Hung.*, 10:299–304, 1959.

[Vid86] E. Vidal. "An Algorithm for Finding Nearest Neighbors in (Approximately) Constant Average Time Complexity". *Patt. Rec. Letters*, 4:145–157, 1986.

[VM06] P. Vary and R. Martin. *Digital Speech Transmission - Enhancement, Coding and Error Concealment*. Wiley, Chichester, 2006.

[Woo69] R. Wood. "On Optimal Quantization". *IEEE Transactions on Information Theory*, IT-5:248–252, mar 1969.

[XIB88] C. Xydeas, M. Ireton, and D. Baghbadrani. "Theory and Real-Tine Implementation of a CELP Coder at 4.8 and 6.0 Kbits/sec using Ternary Code Excitation". *Proc. Int. Conf. on Comrns. IERE Loughborough*, 1988.

[Yeu02] R. W. Yeung. *A First Course in Information Theory*. Springer Verlag, Berlin, 2002.

[Zad66] P. Zador. "Topics in the Asymptotic Quantization of Continuous Random Varianles". Bell Lab. Tech. Memo, 1966.

[ZF99] E. Zwicker and H. Fastl. *Psychoacoustics. Facts and Models*. Springer-Verlag, Berlin, Heidelberg, New York, 2nd edition, 1999.

[ZN77] R. Zelinski and P. Noll. "Adaptive Transform Coding of Speech Signals". *Proc. of Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 299–309, 1977.

[Zwi61] E. Zwicker. "Subdivision of the audible frequency range into critical bands". *The Journal of the Acoustical Society of America*, 33, February 1961.

[Zwi82] E. Zwicker. *Psychoakustik*. Springer Verlag, Berlin, Heidelberg, New York, 1982.

Curriculum Vitae

Angaben zur Person

Name	Hauke Ulrich Krüger
Geburtsort und -tag	Lübeck, 19. Juni 1975
Staatsangehörigkeit	deutsch
Familienstand	verheiratet, zwei Kinder

Schulbildung

1981 – 1985	Gerhard Hauptmann Grundschule, Stockelsdorf.
1985 – 1994	Leibniz Gymnasium, Bad Schwartau.

Studium

Okt. 1994 – Feb. 2000	RWTH Aachen: Studium der Elektrotechnik, Vertiefungsrichtung „Nachrichtentechnik“.
4. Februar 2000	Diplom.

Studienbegleitende Tätigkeiten

Juni 1994 – Sep. 1994	Lubeca/RADE GmbH, Lübeck, Grundpraktikum.
Feb. 1995 – März 1995	Deutsche Telekom, Lübeck, Grundpraktikum.
Jan. 1997 – Juli 1998	Studentische Hilfskraft am Institut für Integrierte Systeme der Signalverarbeitung (ISS) der RWTH Aachen.
Sept. 1998 – Feb. 1999	Conexant Systems, Inc., Newport Beach, CA, USA, Fachpraktikum.

Berufspraxis

März 2000 - Sept. 2000	AXYS Design Automation GmbH in Aachen und Irvine, CA, USA, Tätigkeit als Software-Entwicklungs-Ingenieur.

Wissenschaftliche Tätigkeit

seit Okt. 2000	Wissenschaftlicher Angestellter am Institut für Nachrichtengeräte und Datenverarbeitung (IND) der RWTH Aachen,
1. April 2004	Ernennung zum wissenschaftlicher Assistenten, Forschungsschwerpunkt „Digitale Audiosignalverarbeitung“, insbesondere „Algorithmen-Entwicklung zur Codierung von Sprach- und Audiosignalen“, Nebenschwerpunkt „Entwicklung von Software-Werkzeugen zur schnellen Erstellung von Echtzeit-Prototypen für die Audiosignalverarbeitung“.

Aachen, den 25. Februar 2010