HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electrical and Communications Engineering

Markus Vaalgamaa

Moving average vector quantization in speech coding

This Master's Thesis has been submitted for official examination for the degree of Master of Science in Espoo on January 27, 1999.

Supervisor of the Thesis: Professor Matti Karjalainen

Instructor of the Thesis: Vesa Ruoppila, M.Sc.

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF MASTER'S THESIS

Author:	Markus Vaalgamaa
Thesis name:	Moving average vector quantization in speech coding
Date:	January 27, 1999 Number of pages: 74
Faculty:	Electrical and Communications Engineering
Professorship:	S-89 Acoustics and Audio Signal Processing
Supervisor:	Professor Matti Karjalainen
Instructor:	Vesa Ruoppila, M.Sc., Nokia Research Center

This Master's Thesis studies quantization of spectral parameters in speech coding. An all-pole filter is employed to model short-term spectral information within each speech frame. Line spectral frequency (LSF) representation is used for quantization and interpolation of the filter parameters. The properties of speech spectrum and LSF parameters are discussed thorougly. It is shown that the LSF representation has several properties which makes it suitable for quantization.

Since speech is quasi-stationary, predictive coding methods can exploit the correlation between LSF parameters of adjacent frames. Thus this thesis focuses on quantizer structures whose performance is improved using moving average predictors. Three predictive structures are introduced. A training algorithm for quantizers using these predictors is also presented. The thesis shows that a quantizer using an inter-split predictor obtains a maximal performance gain of one bit per frame compared to the conventional structures. In addition, this quantizer achieves spectral distortion of 1 dB and outlier percentage of 2 % at 23 bits per frame. This performance limit is known as transparent quality in literature.

The perceptual quality of the quantizers is evaluated with subjective listening tests. The quantizers are installed in an IS-641 speech codec. The tests imply that perceptually transparent quality can be achieved even with a 20-bit quantizer in a noiseless environment. Furthermore, the tests show that in unvoiced segments of speech the spectral parameters can be loosely quantized, whereas voiced segments have to be quantized accurately.

Keywords: Vector quantization, moving average prediction, speech coding, speech processing, linear prediction, line spectral frequency, LSF, line spectral pair, LSP

TEKNILLINEN KORKEAKOULU

DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Markus Vaalgamaa
Työn nimi:	Liukuvakeskiarvoinen vektorikvantisointi puheenkoodauksessa
Päivämäärä:	27. 1. 1999 Sivumäärä: 74
Osasto:	Sähkö- ja tietoliikennetekniikan osasto
Professuuri:	Akustiikka ja äänenkäsittelytekniikka, koodi S-89
Työn valvoja:	Professori Matti Karjalainen
Työn ohjaaja:	DI Vesa Ruoppila, Nokia Tutkimuskeskus

Tässä diplomityössä tutkitaan spektrin parametrien kvantisointia puheenkoodauksessa. Nolla-napasuotimella mallinnetaan spektrin lyhytaikaista informaatiota jokaisessa puhekehyksessä. Spektriviivataajuuksiin perustuvaa esitystä (LSF) käytetään suotimen parametrien kvantisointiin sekä interpolointiin. Puheen spektrin ja LSF-parametrien ominaisuuksia esitellään perusteellisesti. Työssä osoitetaan, että LSF-esityksellä on monia ominaisuuksia, jotka tekevät sen sopivaksi kvantisointia varten.

Koska puhe on kvasistationaarista, ennustavat koodausmenetelmät voivat hyödyntää vierekkäisten kehysten LSF-parametrien korrelaatiota. Tämä työ keskittyy kvantisoija-rakenteisiin, joiden suorituskykyä on parannettu liukuvakeskiarvoisilla (MA) prediktoreilla. Kolme prediktiivistä rakennetta esitellään. Näitä ennustajia käyttävien kvantisoijien opetusalgoritmi esitellään. Työ osoittaa, että kvantisoija, joka käyttää splittien välistä ennustajaa, saavuttaa parhaimmillaan suorituskykyedun yksi bitti per kehys verrattuna perinteisiin rakenteisiin. Lisäksi tämä kvantisoija saavuttaa 1 dB spektrivääristymän sekä 2 % määrän arvoille, jotka ylittävät 2 dB, 23 bitillä per kehys. Tämä suorituskykyraja tunnetaan kirjallisuudessa “transparenttina” eli alkuperäistä vastaavana laatuna.

Kvantisoijien perkeptuaalinen laatu arvioidaan subjektiivisilla kuuntelukokeilla. Kvantisoijat on asennettu IS-641 puhekoodekkiin. Testit antavat ymmärtää, että perkeptuaalisesti vastaava laatu voitaisiin saavuttaa jopa 20 bitin kvantisoijalla häiriöttömässä ympäristössä. Lisäksi testit osoittavat, että puheen soinnittomissa osissa spektriparametrit voidaan kvantisoida löysästi, kun taas soinnilliset osat tulee kvantisoida tarkasti.

Avainsanat: Vektorikvantisointi, liukuvakeskiarvoinen ennustaminen, MA ennustaminen, puheenkoodaus, puheenkäsittely, lineaarinen ennustus, spektriviivataajuus, spektriviivapari, viivaspektripari

Preface

First of all, I would like to thank my instructor, Vesa Ruoppila for his continued guidance, support, and admirable patience during my extended work. His experience and theoretical background were fundamentally important for my work described in this thesis. I would also like to thank him as my supervisor at the Nokia Research Center and an excellent coworker and friend.

I would also like to thank my supervisor, Professor Matti Karjalainen, for his support and knowledge and for being an excellent teacher. Without his encouragement and support this thesis would not have been completed within the allowable dime duration.

In terms of financial support and coordination of the work, I would like to thank Professor Petri Haavisto and Jari Hagqvist, M.Sc. at the Nokia Research Center. In addition, I am grateful for the Speech and Audio Systems Laboratory at Nokia Research Center, Tampere, for the appropriate working environment.

Moreover, I would like to thank all former and present colleagues at Nokia Research Center and at the Laboratory of Acoustics and Signal Processing, Helsinki University of Technology for creating a friendly and stimulating atmosphere and providing fruitful ideas and support during the work. Especially, I would like to mention Ari Heikkinen, M.Sc. and Samuli Pietilä, Lic.Tech. at the Nokia Research Center.

My special thanks go to my parents, sister and brothers, for giving me encouragement and amusement during the writing of this thesis. And finally, I express my sincere gratitude to my beloved Sanna for her warmth, support, and patience.

Espoo, January 26, 1999

Markus Vaalgamaa

Lauttasaarentie 38 A 9
00200 Helsinki
Finland

Tel: +358 50 588 1188
email: markus.vaalgamaa@hut.fi

Table of Contents

List of Symbols

Symbol	Description
a_i	An *i*-th LPC coefficient
$a_1, \ldots, a_p$	LPC coefficients
$A(z)$	A linear prediction filter, known as the "whitening" filter
$A_i(z)$	An original LPC polynomial of the *i*-th frame
$\hat{A}_i(z)$	A quantized LPC polynomial of the *i*-th frame
b_{ijk}	An element of the *i*-th prediction matrix
$\mathbf{B}_i$	A prediction coefficient matrix of an order *i*
$\mathbf{B}_{ijk}$	A sub-matrix of prediction coefficient matrix of an order *i*
$c(n)$	An *n*-th elements of fixed codebook vector of the ACELP synthesis model
$\mathbf{C}$	A codebook of a quantizer
$\mathbf{C}_k$	A codebook of the *k*-th split
$d(\ )$	Any distortion measure
d_{AQ}	An average quantization error measure
d_{AQi}	An average quantization error of the *k*-th split
d_{ED}	A weighted Euclidean distance measure
d_i	A distance between (*i*–1)-th and (*i*+1)-th LSF parameter
DMOS	Decredation Mean Opinion Score of listening test
$\mathbf{e}(t)$	A prediction error vector at time instant *t*
f_{bwe}	A bandwidth expansion, which has a fixed value 60 Hz
f_h	An upper frequency (Hz) limits for integration of spectral distortion
f_l	A lower frequency (Hz) limits for integration of spectral distortion
F_s	A sampling frequency, which has a fixed value 8000 Hz

g_c	A fixed codebook gain of the ACELP synthesis model
g_p	A pitch or adaptive codebook gain of the ACELP synthesis model
$H(z)$	A transfer function of a filter *A(z)*
i	A running index
$\mathbf{I}$	An identity matrix
j	A running index
$J(\ )$	An objective function for minimization in prediction paremeter estimation
k	A running index
K	A number of speech segments
L	A number of samples in the speech segment
m_j	A size of some matrix
M	A number of training vectors
n	A running index, used as a discrete time instant
n_B	An order of a moving average predictor
N	A number of codevectors in a codebook
N_k	A number of codevectors in the *k*-th split
$N_{\text{bit}k}$	A bit rate of the sub-quantizer of the split *k* per frame, defined as $\log_2(N_k)$
N_{bit}	An overall bit rate of the quantizer per frame i.e. a sum of $N_{\text{bit}k}$:s
p	An order of a linear predictive filter
P	An optimal partition of an encoder
$P(z)$	A symmetric polomial used to solve odd LSF parameters
$P_i(f)$	An original LP power spectra of the *i*-th frame
$\hat{P}_i(f)$	A quantized LP power spectra of the *i*-th frame
$\mathbf{q}$	A selector vector for choosing free parameters from the parameter vector $\boldsymbol{\theta}$
q_i	An *i*-th element of the selector vector $\mathbf{q}$
$Q(z)$	An antisymmetric polynomial used to solve even LSF parameters
$\mathbf{r}$	A vector of minimization of objective function $J(\ )$ in respect to $\boldsymbol{\theta}'$
$r(i)$	An *i*-th autocorrelation coefficient
$r'(i)$	An *i*-th modified autocorrelation coefficient
R_i	An *i*-th optimal partition cell a codebook
$\mathbf{R}^{-1}$	An inverse matrix of minimization of objective function $J(\ )$ in respect to $\boldsymbol{\theta}'$
Rel. bit rate	A relative bit rate by comparing a quantizer to another
s	A number of splits in the quantizer

$s(n)$	A speech sample at discrete time instant n
$\hat{s}(n)$	A reconstructed speech of the ACELP synth. model at discrete time instant n
$\hat{s}'(n)$	A synthesis speech of the ACELP synthesis model at discrete time instant n
segSNR	A segmental signal-to-noise ratio
SD_i	A spectral distortion of the i-th frame presented in unit dB
$\mathrm{SD}_{\mathrm{ave}}$	An average spectral distortion presented in unit dB
$\mathrm{SD}_{2\mathrm{dB}}$	A percentage (%) of outlier frames having spectral distortion over 2 dB
$\mathrm{SD}_{4\mathrm{dB}}$	A percentage (%) of outlier frames having spectral distortion over 4 dB
t	A time instant, corresponds to the t-th frame
$u(n)$	An n-th elements of total excitation of the ACELP synthesis model
$\mathbf{u}(t)$	A chosen codevector from a codebook $\mathbf{C}$ at time instant t
$\mathbf{u}_k(t)$	A chosen codevector from a codebook $\mathbf{C}_k$ of the k-th split at time instant t
$\mathbf{u}_k^{(i)}$	An i-th codevector of a codebook $\mathbf{C}_k$ of the k-th split
Unst.	A number of frames having an unstable filter
$v(n)$	An n-th elements of adaptive codebook vector of the ACELP synth. model
w_i	A square root of i-th diagonal element of the weighting matrix $\mathbf{W}$
$w_{lag}(i)$	An i-th multiplication factor of bandwidth expansion
$\mathbf{W}$	A weighting matrix of Euclidean distance measure
$\mathbf{x}$	An original LSF vector
$\hat{\mathbf{x}}$	A quantized LSF vector
$\hat{\mathbf{x}}(t)$	A quantized LSF vector at time instant t
$\hat{\mathbf{x}}_k(t)$	A quantized LSF vector of the k-th split at time instant t
x_i	An i-th LSF parameter
$\mathbf{y}_i$	A i-th codevector of a codebook $\mathbf{C}$
z	A complex variable of z-transform
$\boldsymbol{\theta}$	A parameter vector containing prediction coefficients
$\boldsymbol{\theta}'$	A modified parameter vector containing free elements of $\boldsymbol{\theta}$
θ_i	An i-th element of the parameter vector $\boldsymbol{\theta}$
$\boldsymbol{\varphi}(t)$	A vector containing the codevectors $\mathbf{u}(t), \ldots, \mathbf{u}(t-n_B)$
$\Phi(t)$	A matrix containing $\boldsymbol{\varphi}(t)$ which is multiplied with the Kronecker product
$\Phi'(t)$	A modified matrix from $\Phi(t)$ containing its free columnvectors
ω_i	An i-th line spectral frequency (LSF) presented in unit rad/s

List of Abbreviations

Abbreviations	Description
(i, j)	A codec using i excitation pulses and j bits for LSF quantization
ACELP	Algebraic code excited linear prediction
ADPCM	Adaptive Differential Pulse Code Modulation
AR	Autoregressive
CELP	Code excited linear prediction
DCR	Degradation Category Rating listening test
DMOS	Degradation Mean Opinion Score, a measure of DCR listening test
DP-SVQ	MA-SVQ using a diagonal matrix predictor
$DP_{IS\text{-}641}$-SVQ	The LSF quantizer of IS-641 speech codec
n_B-DP-s-SVQ	MA-SVQ using a diagonal matrix predictor, the integer s is optional
Fi	The i-th female voice in listening tests
FP-SVQ	MA-SVQ using a full matrix predictor
n_B-FP-s-SVQ	MA-SVQ using a full matrix predictor, the integer s is optional
GLA	The Generalized Lloyd algorithm
IPA	International Phonetic Alphabet
IRS	Intermediate reference system
IS-SVQ	MA-SVQ using a inter-split predictor
(IS)	The original IS-641 speech codec in listening tests
IS-641	Enhanced full rate speech codec of US-TDMA digital cellular system
n_B-IS-s-SVQ	MA-SVQ using a inter-split predictor, the integer s is optional
n_B-$IS_{12\ldots s}$-SVQ	IS-SVQ, using ascending quantization order of splits
n_B-$IS_{s(s\text{-}1)\ldots1}$-SVQ	IS-SVQ, using declining quantization order of splits

ITU	International Telecommunication Union
ITU-T	ITU, Telecommunication standardization sector
LBG	An algorithm for vector quantizer design, by Linde, Buzo, and Gray
LP	Linear prediction
LPC	Linear predictive coding
LSP	Line spectral pair
LSF	Line spectral frequency
M*i*	The *i*-th male voice in listening tests
MA	Moving average
MA-SVQ	A moving average split vector quantizer
MNRU	Modulate Noise Reference Unit, a reference sample in DCR test
(NQ)	IS-641 where LSF parameters are not quantized, in listening tests
NTT	Nippon Telephone and Telegraph Corporation
PCM	Pulse Code Modulation
SD	Spectral distortion measure
segSNR	Segmental Signal-to-Noise Ratio
SNR	Signal-to-Noise Ratio
TIA	Telecommunications Industry Association
TIMIT	TIMIT speech database
TDMA	Time Division Multiple Access network
(u)	IS-641 where LSF parameters are quantized in unvoiced frames
US-TDMA	North American Time Division Multiple Access network
(v)	IS-641 where LSF parameters are quantized in voiced frames
VQ	Vector quantization

1 Introduction

Speech coding has become one of the most important areas of modern digital communication during last two decades. Digital mobile telephone plays an important role in every-day life for millions of people. Transmission and storage of speech and audio signals are in enormous growth because of Internet communication. Multimedia applications apply more and more audio and speech. The development of microprocessors and signal processing hardware stimulates new ideas to use speech processing. For all these reasons the demand for faster, more efficient, more reliable, and better quality systems is continually expanding.

In speech coding the primary goal is to achieve high perceived quality of reconstructed speech signal at low cost. The costs are composed of several issues, such as bit rate, complexity, and robustness to transmission errors. The weight of these issues depends on application, although the bit rate or compression rate seems to have a substantial importance especially in mobile communication.

In digital speech communications, speech is generally bandlimited below 4 kHz (or 3.2 kHz) and sampled at 8 kHz. Typically speech samples are amplitude quantized to 8–16 bits. The quantization can be either uniform or nonuniform. Nonuniform quantization can be used at lower bit rates since human hearing sensitivity is logarithmic [1]. Typical examples of nonuniform quantization are A-law companding used in the European telecommunications systems and μ-law companding used in the American and Japanese telecommunication systems [2]. The simplest coding technique is Pulse Code Modulation (PCM) which is simply a quantizer of isolated sample amplitudes. Speech coded at 64 kbit/s using logarithmic PCM is considered as “non-compressed” and is often used as a

reference for comparisons. An advanced conventional coding technique is Adaptive Differential Pulse Code Modulation (ADPCM) operating at 32 kbit/s. The perceived quality of these coding schemes is referred often as *toll quality* or *telephone quality*.

Sophisticated speech coding methods that reduce redundancy and remove perceptually irrelevant information in speech have enabled to achieve high quality at lower bit rates. At rates between 16 kbit/s and 32 kbit/s *linear predictive coding* (LPC) is typically used to model the speech signal. Further *linear-prediction-based analysis-by-synthesis* coding can be used to increase the efficiency of quantizing the speech signal at coding rates between 4 kbit/s and 16 kbit/s. A popular quantization scheme today, *Code Excited Linear Prediction* (CELP) is based on analysis-by-synthesis coding. CELP coders employ vector codebooks to code excitation signal. More extensive discussion on speech coding algorithms can be found in [3, 4].

One key factor in this progress is the rapid development of signal compression techniques. These techniques can be either *lossless* or *lossy*. In lossless coding, a signal can be perfectly reconstructed. However, *a compression ratio* achieved by lossless compression is slight for current demands. Thus lossy compression techniques are typically used. The objective of lossy compression is to minimize the distortion between the original and the reconstructed signal.

Vector quantization (VQ) is one of the most powerful lossy coding methods. One of the most often used applications for VQ is the quantization of speech spectrum. The matter has been intensively studied and developed during the last decades. VQ has proven to be efficient for coding of LPC parameters and thus it is widely used in modern speech coders. Despite of the progress, the transmission of the spectrum parameters requires between 1 and 2 kbit/s, which is a major contribution to the overall bit rate for low-rate speech codecs. Therefore, it is important to enhance the quantization methods. The *predictive quantization* is shown to be promising method.

The LP analysis for modeling the speech spectrum is presented in Section 1.1. A typical environment for this analysis is given in Section 1.2, where the IS-641 speech codec is presented. The section focuses on the LP analysis of the codec. The organization of this thesis is outlined in Section 1.3. The section also gives the motivation for the quantization of speech spectrum.

1.1 Linear predictive speech coding

Most modern speech coders are based on the source-filter model of human speech production. In such coders a synthesis filter, which roughly models the human vocal tract, is driven by an excitation signal which essentially models the flow of air through the vocal

chords. The synthesis filter can be modeled using linear prediction. The rate at which the shape of vocal tract changes is limited, and typically an update rate of 50 Hz is sufficient for the model. Thus the LP analysis is typically done once in a 20 ms frame.

In linear predictive coding analysis it is assumed that the current speech sample can approximately be predicted by a linear combination of *p* past samples; that is

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k \, s(n-k), \tag{1.1}$$

where *s(n)* is the speech sample at discrete time instant *n*, *p* is the order of analysis and a_1, …, a_p are the LPC coefficients. The order *p* of the system is chosen such that the estimate of the spectral envelope is adequate. A common rule of thumb is to allow one pole pair for every formant present in the spectrum. For a speech signal sampled at 8 kHz, the value of *p* is typically ten.

The transfer function *H(z)* of the linear prediction speech model is

$$H(z) = \frac{1}{A(z)}, \tag{1.2}$$

where the filter *A(z)* is known as the “whitening” filter or “inverse” filter of *H(z)*, defined as

$$A(z) = 1 + \sum_{k=1}^{p} a_k \, z^{-k} \, . \tag{1.3}$$

H(z) is also referred to as an *all-pole* model of the speech signal.

LPC coefficients can be solved using for example the autocorrelation method or the covariance method. The autocorrelation method produces an autocorrelation matrix which has a Toeplitz structure and thus the LPC coefficients can be solved through computationally fast algorithms such as the Levinson-Durbin algorithm [5, 6].

1.2 IS-641 speech codec

The IS-641 [7] was chosen as a base for this study since it is a modern, good quality and low bit rate codec. In 1996, the enhanced full rate speech codec IS-641 has been standardized for the US-TDMA digital cellular system IS-136. The predecessor of IS-641, the full rate codec of the North America TDMA employs 7.95 kbit/s for speech coding and

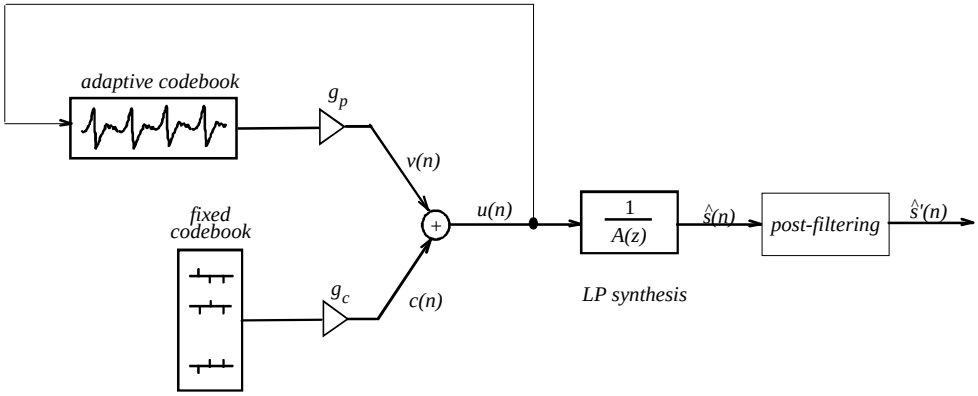

Figure 1.1. Simplified block diagram of the ACELP synthesis model.

5.05 kbit/s for error protection resulting in 13.0 kbit/s in total. Because of the rapid advances in speech coding technology, Telecommunications Industry Association (TIA) considered the developing of improved full rate codec for the TDMA system. The bit rate of the speech codec was reduced to 7.4 kbit/s and the error protection was increased to 5.6 kbit/s. The codec developed jointly by Nokia and University of Sherbrooke was approved as the IS-641 standard [8]. Due to the improved speech coding algorithms, the IS-641 speech codec improves significantly the speech quality compared to the full rate codec of IS-136.

The IS-641 speech codec is based on the *Algebraic Code Excited Linear Prediction* (ACELP) algorithm [3]. The main elements of ACELP are a linear prediction synthesis filter, a long-term predictor and a pulse excitation generator. The ACELP speech synthesis model is shown in Figure 1.1. In this model, the input signal of the LP synthesis filter is constructed by adding two excitation vectors from adaptive and fixed codebooks. The adaptive codebook is used to generate the pitch to the excitation signal. The fixed codebook generates the excitation signal using multi-pulse permutations. The speech is synthesized by feeding the two properly chosen vectors from these codebooks through the synthesis filter.

The IS-641 codec operates on speech frames of 20 ms corresponding to 160 samples at the sampling frequency of 8000 Hz. The speech frame is divided into four subframes of 5 ms each. In each frame the speech signal is analyzed by the encoder. The encoder extracts the parameters to represent the speech frame and packs them to a bitstream. The extracted parameters are linear prediction filter coefficients, and indices and gains for the adaptive and the fixed codebooks. Consequently the decoder unpacks the bitstream and reconstructs the synthesis speech signal. The following section concentrates only on the front end of the IS-641 codec. The detailed description of the encoder and decoder can be found in [7].

1.2.1 Linear prediction analysis of IS-641

Since this thesis concentrates on the quantization of LP parameters, the front end of the IS-641 codec is presented here in detail. The front end is used to produce the spectral parameters of training and validation databases for the design and evaluation of vector quantizers examined in this thesis.

First the input speech signal is high-pass filtered with a cut-off frequency of 80 Hz. The LP analysis is carried out for once in frame of 20 ms. Autocorrelation coefficients are calculated using a 30 ms asymmetric *analysis window*, presented in Figure 1.2. A lookahead of 5 ms is used in the analysis window. The window has its weight concentrated on the fourth subframe and it consists of two parts: the first part is a half of a Hamming window and the second half is a quarter of a cosine function cycle. The shape of the window is attributable to interpolation of LP filters described later in this section.

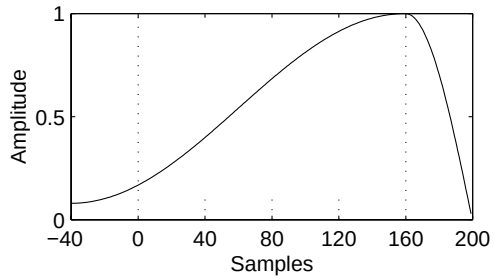

Figure 1.2. Asymmetric analysis window in IS-641.
Long and short dotted lines show the frame and subframe boundaries, respectively.

Since the LP analysis may generate synthesis filters with sharp spectral peaks, *bandwidth expansion* is employed. The expansion affects especially the formant peaks in the magnitude response of the filter. Typically the bandwidth expansion is used to avoid unnatural synthesized speech of high-pitched voices, when LP analysis has problems in estimating the spectral envelope [9]. In addition, the expansion increases the robustness of filter against quantization errors.

Bandwidth expansion of 60 Hz is done by lag windowing the autocorrelation coefficients [10]. The autocorrelation coefficients are multiplied by factors

$$w_{lag}(i) = e^{-\frac{1}{2}\left(\frac{2\pi f_{\text{bwe}} i}{F_s}\right)^2}, \quad i = 1, \ldots, 10, \tag{1.4}$$

where f_{bwe} = 60 Hz is the bandwidth expansion and F_s = 8000 Hz is the sampling frequency. Thus modified autocorrelations are $r'(i) = w_{lag}(i)r(i)$ for i = 1, …, 10.

Further so-called *white-noise correction* is made to reduce numerical problems of LP analysis. Since speech has a strong low-pass filtered spectrum (-6 dB/octave), the spectrum shows a large dynamic range. Although bandwidth expansion minimizes the dynamic range of spectrum by reducing its peaks, the high frequency components in the speech spectrum have very low amplitude. The correction is used since the LP analysis requires high computational precision to capture the description of features at the high end of the speech spectrum. More importantly, when these features are very small, the autocorrelation matrix can become singular, resulting in computational problems. By adding to the signal a low-level noise, the dynamic range of power spectrum is reduced and the numerical problems can be avoided. Thus the first autocorrelation coefficient is multiplied by the white noise correction factor 1.0001, $r'(0) = 1.0001r(0)$, which is equivalent to adding a noise floor at –40 dB to the signal.

The LP coefficients are solved from the modified autocorrelations using the Levinson-Durbin algorithm. The coefficients are then converted into *line spectral frequency* (LSF) parameters [11] for quantization and interpolation purposes. The LSF parameters are quantized using *split vector quantization* (SVQ). A first order moving average predictor is applied to the LSF vector. The LSF residual vector is split into three subvectors of dimensions 3, 3 and 4. These subvectors are quantized with 8, 8 and 9 bits resulting in 26 bits for each 20 ms frame. Thus the quantization of LP parameters requires 1300 bit/s that is nearly 20 % of the overall bit rate of the codec.

The set of quantized and unquantized LSF parameters is used to construct the filters for the fourth subframe. However the first, second, and third subframes use interpolated filters from the spectral parameters of the current and the previous frame. The linear interpolation is done using cosines of the LSF parameters. The cosines are used since they allow to make the interpolation with smaller amount of calculations and still offer similar results than interpolation of the LSF parameters. The interpolated quantized and unquantized filters are converted back to LP filter coefficients to construct the filters for the subframes.

1.3 Organization of the Thesis

The intent of this thesis is to examine effective methods for linear predictive coding of spectral parameters. The predictors of different structures are studied and the performance of quantizers using these predictors are evaluated. The novel prediction scheme, called *inter-split prediction* is introduced for split vector quantization. The lowest bit rate of the quantizers achieving transparent quality is examined. In addition the subjective quality of quantizers is studied.

Chapter 2 presents the LSF representation of LPC parameters. It describes properties of the representation in the context of spectrum quantization. The training and the validation

databases of this thesis are presented. Objective distortion measures that evaluate coding performance are defined in the end of the chapter. Chapter 3 provides an overview of linear predictive vector quantization techniques. The quantizers using moving average predictors are introduced. A training algorithm of such quantizers is presented. Chapter 4 presents objective results of the quantizers. The efficiency of the quantization structures is evaluated and lowest bit rates of quantizers achieving the transparent quality are examined.

The lack of exact objective measures of speech quality makes speech coding a particularly challenging task. Therefore, subjective quality of the quantizers is studied in Chapter 5. The lowest bit rate of quantization achieving the transparent quality in IS-641 speech codec is examined. In addition the relation of the subjective measures compared to the objective measures is analyzed. Chapter 6 concludes the thesis with a summary and suggestions for future investigation.

2 Line Spectral Frequencies

Since the beginning of speech coding several quantization methods have been employed for the LP filter. It has been noticed that there are several requirements for desirable representation. Firstly, it is necessary that an all-pole filter remains stable after the quantization. Secondly, the representation should be reversible such that the original filter can be recovered from the transformation. Thirdly, a small error in the parameters shall correspond to a small deviation of the LP power spectrum, i.e. the parameters shall have a proper spectral sensivity. Finally the parameters should be suitable for quantization in the context of used measures.

In the literature, a number of such representations have been proposed. One of the first studies on alternative representations for the LP filter was performed by Gray *et al.* in 1977 [12]. In that paper a comparison between reflection coefficient, log area ratio, and inverse sine (arcsine) of reflection coefficient representations was performed. They showed that log area ratios and inverse sine of reflection coefficients were equally good for quantization whereas performance of reflection coefficients was slightly poorer.

Since 80s the line spectral frequency (LSF) representation has become the dominant parametrization for the quantization of LPC parameters. Firstly the representation fulfills the previously presented requirements and moreover it is shown to be suitable for quantization. Authors of [13], [14] and [15] have reported that the LSF representation reduces the bit rate 25–30 % compared to log area ratio representation. One of the best comparisons of the alternative representations is presented in Paliwal and Kleijn [16]. They have demonstrated the advantages of the LSF representation over the reflection coefficient, log area ratio and inverse sine (arcsine) of reflection coefficient representations both in

scalar and vector quantization. They have also shown that the LSF representation is also the best of known methods for interpolation of LPC parameters.

This chapter presents the LSF representation of LPC coefficients and introduces its basic properties. Speech databases for training and validation of the LSF quantizers are presented and the statistics of the training database are shown. The chapter also demonstrates the reasons for the success of the LSF representation with examples. Moreover, the objective distortion measures used in this thesis, such as spectral distortion measure, weighted Euclidean distortion measure, average quantization error and segmental signal-to-noise ratio, are presented.

2.1 LSF representation

The line spectral frequency representation was introduced by Itakura [11] as an alternative parametric representation of linear prediction coefficients. The LSF representation, also known as line spectrum pair (LSP) representation, has a number of properties, including a bounded range, a sequential ordering of the parameters and a simple check for the filter stability, which makes it desirable for quantization. In addition, the LSF representation is a frequency domain representation and, hence, can be used to exploit properties of the human perception system.

The inverse LP filter defined in Section 1.1 is given by

$$A(z) = 1 + a_1 z^{-1} + \cdots + a_p z^{-p}, \tag{2.1}$$

where p is the order of the filter and a_i is the ith coefficient of the filter. Typically $p = 10$ in speech coding.

To define the LSF parameters, polynomial *A(z)* is used to construct two polynomials:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}), \tag{2.2}$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}). \tag{2.3}$$

The polynomial *P(z)* is symmetric and the polynomial *Q(z)* is antisymmetric. Soong and Juang [13] have shown that if *A(z)* is minimum phase, then the zeros of *P(z)* and *Q(z)* are on the unit circle and they are interlaced with each other. Therefore *P(z)* and *Q(z)* can be factored as follows:

$$P(z) = \begin{cases} (1 + z^{-1}) \prod_{i=1,3,\ldots,p-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}) & p \text{ even} \\ \prod_{i=1,3,\ldots,p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) & p \text{ odd} \end{cases} \quad (2.4)$$

$$Q(z) = \begin{cases} (1 - z^{-1}) \prod_{i=2,4,\ldots,p} (1 - 2z^{-1}\cos\omega_i + z^{-2}) & p \text{ even} \\ (1 - z^{-2}) \prod_{i=2,4,\ldots,p-1} (1 - 2z^{-1}\cos\omega_i + z^{-2}) & p \text{ odd} \end{cases} \quad (2.5)$$

where ω_1, ω_2, ..., ω_p are the phase angles of the zeros of the polynomials, such that

$$0 < \omega_1 < \omega_2 < \cdots < \omega_p < \pi. \quad (2.6)$$

The phase angles ω_1, ω_2, ..., ω_p are called the line spectral frequencies of *A(z)*. The ascending order of LSF parameters ensures the stability of the LP synthesis filter, which is an important pre-requirement for speech coding applications. The LSF parameters may be calculated from Equation (2.2) and Equation (2.3) using several methods. Soong and Juang [17] compute the LSF parameters applying a discrete cosine transformation and Kabal and Ramachandran [18] use Chebyshev polynomials.

The transformation from LPC coefficients to the LSF parameters is reversible and *A(z)* can be obtained from Equation (2.4) and Equation (2.5) as

$$A(z) = \frac{1}{2}\left[P(z) + Q(z) \right]. \quad (2.7)$$

2.2 Training and testing databases

In this work, two separate databases are used for training and validating of line spectrum frequency vector quantizers. The databases are completely separate, and do not contain common sentences or speakers. This is a result of a common approach in which the quantizer is desired to concentrate only on the main features of the database and not to learn the material in detail. The databases contain sentences of different languages and dialogues spoken by native speakers. Both the training and the validation databases have been recorded with a high-quality microphone in a quiet room, where room noise level and reverberation are minimized, i.e. noise level is below 30 dBA, without dominant peaks a spectrum and reverberation time is less than 500 ms.

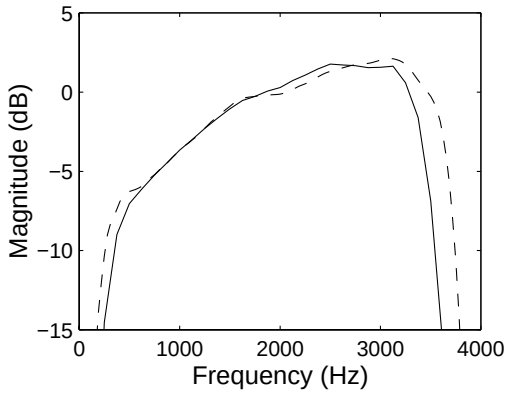

Figure 2.1. The magnitude responses of the IRS filter (flat), solid line, and the modified IRS filter, dotted line.

The speech material in the testing and validation databases has to be prefiltered, before using the speech codec to evaluate LP parameters. The prefiltering has to be made, since a typical telephone line, consisting of everything from a microphone of a telephone to a transmission path, is far from an ideal flat response system. The magnitude response of the telephone line can be simulated with different pre-processing filters. One of the first proposed prefilters is an Intermediate reference system (IRS) filter, which was presented in a beginning of the 1970s. The IRS filter, specified in ITU-T Recommendation P.48 [19], simulates speech signals obtained from an average analog telephone handset. The filter emphasizes the middle frequencies, and its –10 dB attenuation limits are 340 Hz and 3550 Hz. A modified IRS filter has been introduced in ITU-T Recommendation P.830 [20] to correspond to better magnitude characteristics of the modern digital telecommunication systems. Compared to the IRS filtering, the modified IRS filtering provides wider magnitude response, with –10 dB attenuation limits 260 Hz and 3750 Hz. The magnitude responses of the IRS and modified IRS filters are presented in Figure 2.1. Beside these filterings, also clean speech, without pre-processing, is used to simulate flat magnitude responses of a high-quality telephone handset and computer application, where speech is not transmitted through telephone network.

The training database consists of 51 minutes of speech. The source material originates from NTT's *Multi-Lingual Speech Database for Telephonemetry 1994* speech database [21]. The source data, 25 minutes 30 seconds of speech, was low-pass filtered and downsampled to 8000 Hz. Source data was doubled, so that the first half is unfiltered (flat) and the another half is pre-processed with the modified IRS filter. Roughly half of the database consists of American English sentences, spoken by five females, five males and two children. The rest of the material is multi-lingual, comprising Chinese, French, German, Italian, Russian and Spanish. The test database produces a total of 151962 frames.

The validation set consists of a speech sample of 22821 frames in length of 7 min 36 s. The speech sample originates from the TIMIT database [22]. Speech is low-pass filtered and downsampled with a sampling rate of 8000 Hz. The validation data is IRS filtered. It

contains eight American English female voices and twelve male voices. It is worth to point out that modified IRS filtered and unfiltered speech are used in the training database where as conventional IRS filtered speech is used in the validation database. The mismatching filters simulate the real life environment, where the data is not similar to that used in training.

2.3 Properties of LSF representation

In this section, two examples are presented to illustrate the properties of LSF representation in speech coding. Spectrogram of a speech sample is shown and features of the speech spectrum are discussed. To understand the use of LSF parameters with predictive vector quantization, the knowledge of speech itself is inevitable. In the end of this section the localized spectral property of the representation is shown.

As an example of the behavior of the LSF parameters, an English sentence “A major breakthrough has been made.” is visualized in Figures 2.2–2.4. The sentence has been spoken by a native British male, from NTT’s database [21]. The front end of the IS-641 speech codec is applied to calculate the LSF parameters. The sentence consists of 98 frames, each producing one vector of ten LSF parameters. These LSF parameters are used for the fourth subframes whereas the first, second, and third subframes use linearly interpolated LSF parameters, as was discussed in Section 1.2.1.

Each figure has been divided into three graphs. In the top graph the signal is presented. In addition, syllables of words are labeled into the graph. The middle and the bottom graph present a spectrogram of the signal. The spectrogram is calculated by summing the LP power spectrum and the energy of the signal in each subframe. The interpolation of LSF parameters are used to build the filters for the first, the second, and the third frame. In the middle graph the corresponding LSF parameters of filters are presented. The bottom graph visualizes the three-dimensional spectrogram. The resolution of the middle and the bottom graphs is 5 ms which corresponds the subframe length.

It can be seen from the middle and bottom graphs that a cluster of line spectral frequencies (from two to three LSF parameters) characterizes a formant frequency and the bandwidth of a formant depends on the closeness of the corresponding LSF parameters. For example, the formants of phoneme [ei] (using IPA alphabet) in the word ‘major’ can be easily detected during the time interval 0.26 s and 0.40 s (Figure 2.2). The first formant, built up by line spectral frequencies one to three, lies near 400 Hz. The second, the third, and the fourth formant, built up by line spectral frequencies from five to ten, lie near 2000 Hz, 2600 Hz and 3300 Hz, respectively. On the other hand, isolated LSF parameters affect the spectral tilt, for example LSF parameters three to five between 500 Hz and 2000 Hz in phoneme [ei].

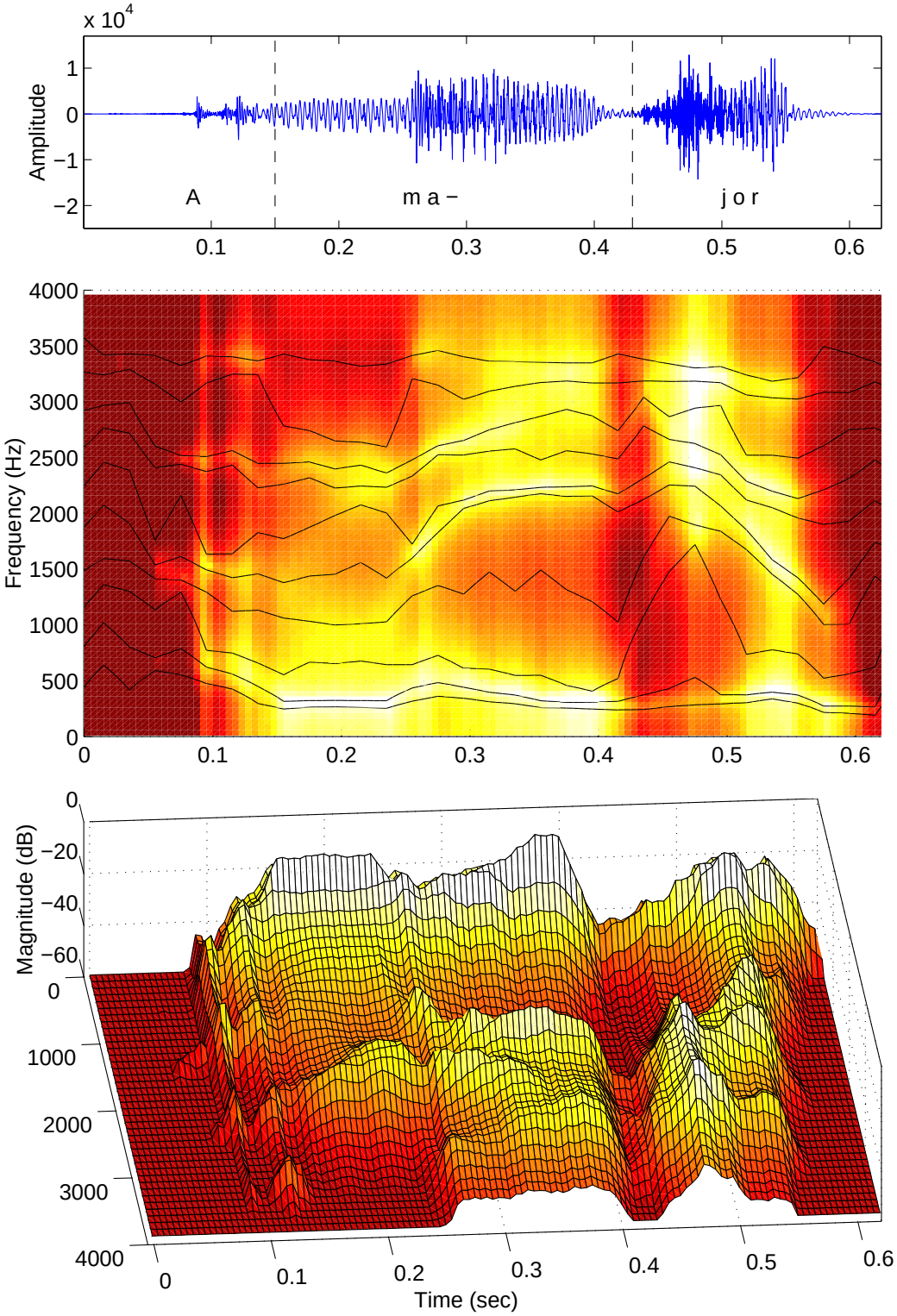

Figure 2.2. The first part of English sentence “A major breakthrough has been made.” spoken by a British male. The signal is presented in the top graph and the borders of syllables are marked with dashed lines. The spectrogram with LSF parameters is in the middle graph, and the bottom graph illustrates the three-dimensional spectrogram of the signal: the brighter color indicates stronger magnitude.

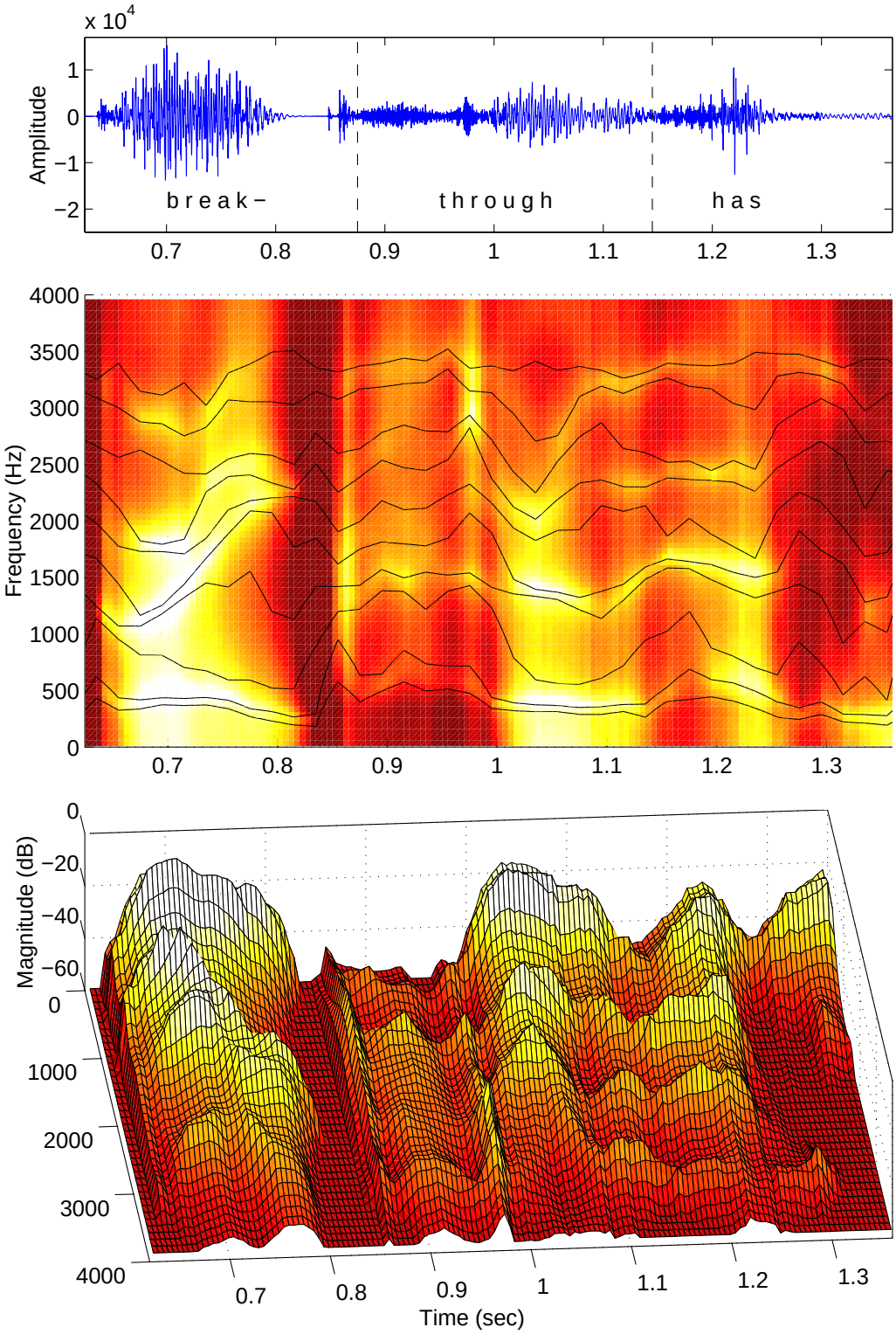

Figure 2.3. The second part of the sentence “A major breakthrough has been made.”

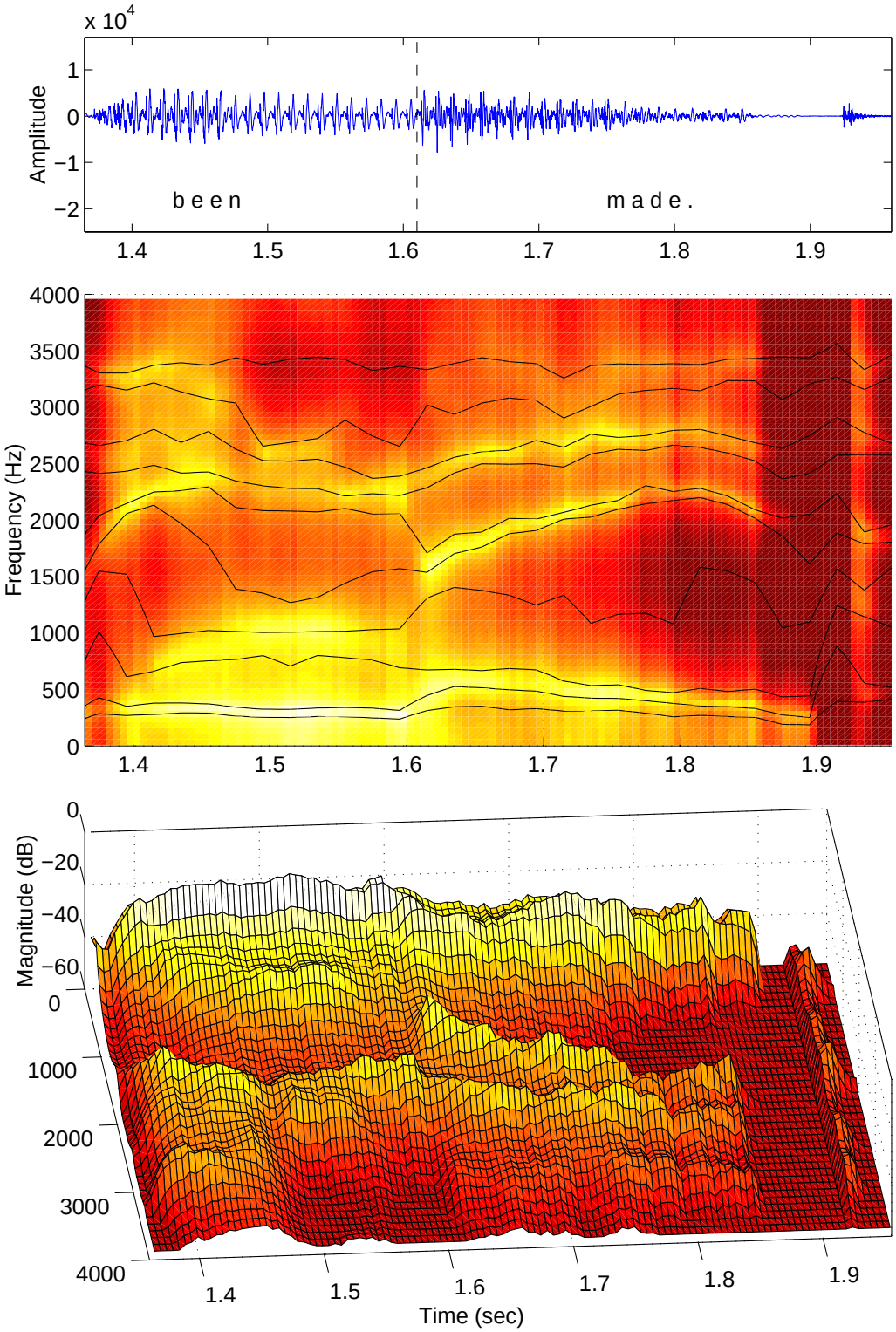

Figure 2.4. The third part of the sentence “A major breakthrough has been made.”

Another noteworthy detail is that during voiced phonemes LSF parameters vary only slightly, see for example the phoneme [ei] in the word ‘major’ or a phoneme [i:] in the word ‘been’. However, during unvoiced phonemes, e.g. a phoneme [z] of the fricative ‘s’ between 1.23 s and 1.30 s, the LSF parameters may change rapidly. Furthermore, during shift to or from unvoiced phoneme there can be considerable changes in spectrum, see for example the change from ‘ma’ to ‘jor’ around 0.4 ms. Also during silences between phonemes the LSF parameters vary considerably. Altogether, the speech spectrum and corresponding LSF parameters have considerable redundant parts, together with rapidly changing and unpredictable parts. Especially lowest LSF parameters are nearly immovable for hundreds of milliseconds and then they can suddenly change to another states. However the previously presented conclusions are not always as evident. It has to be kept on mind that these conclusions apply only in a noiseless environment.

The LSF representation is useful for the interpolation of LP filters. From Figures 2.2–2.4 it can be seen that the frequency and the magnitude of formants behave naturally during interpolation. In addition, the interpolation of LSF parameters of two stable filters produces always a stable filter.

Finally the spectral sensitivities of LSF parameters are localized. Figure 2.5 shows the magnitude response and the LSF parameters of the vowel /a/ from the word ‘major‘. Figure 2.6 illustrates two examples of the same magnitude response in which a single line spectrum frequency is slightly modified. It can be seen that a change in this frequency produces a change in the magnitude response only in its neighboring frequencies. The change of the fourth LSF parameter affects the magnitude response mostly near 1000–1500 Hz. Similarly, a change of the ninth LSF parameter modifies the fourth formant near 3400 Hz. However if the change in a single parameter is large compared to the distance of the neighboring parameters, the entire magnitude response will be damaged.

The localized spectral sensivity property of the LSF parameters has several advantages. Firstly, the LSF representation tolerates small errors in the LSF parameters, i.e. small errors have small influence to the magnitude response of the corresponding LP filter. Secondly, the individual parts of an LSF vector can be independently quantized without leakage of quantization distortion from one spectral region to another. Thirdly, different weights can be given to different LSF parameters with respect to their importance of speech spectrum. Usually the formant areas are perceptually more important than spectral valleys and therefore more weight can be given to the line spectrum frequencies near formants. Furthermore, the human ear cannot resolve differences at high frequencies as accurately as at low frequencies, and thus higher frequency LSF parameters can be quantized inaccurately than lower ones. It is noteworthy that several other LPC representations, such as linear area ratios or reflection coefficients, do not have localized spectral sensitivity and the advantages previously mentioned.

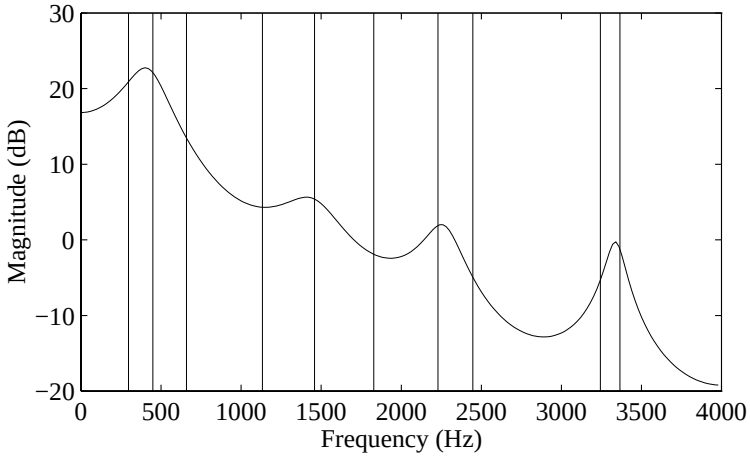

Figure 2.5. The LP spectrum and the LSF parameters of vowel /a/ from the example sentence. The LSF parameters are 297, 449, 659, 1133, 1459, 1828, 2229, 2447, 3243 and 3366 Hz.

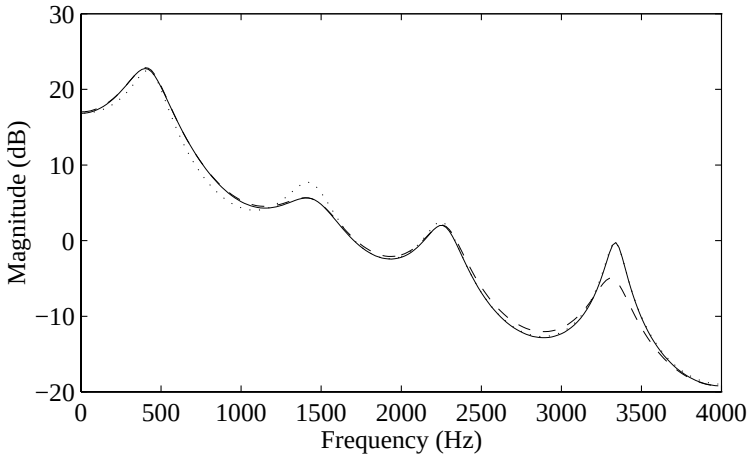

Figure 2.6. Effect of changing values of LSF parameters on the LP power spectrum. The original spectrum is shown by solid line. The 70 Hz change of the fourth LSF from 1130 Hz to 1200 Hz is shown by dotted line and the 120 Hz change of the ninth LSF from 3243 Hz to 3163 Hz is shown by dashed line.

2.4 Statistical properties of training database

To understand the statistical behavior of the speech spectrum and the line spectral frequencies, the training database is studied more accurately. The LSF parameters of the training database are calculated using the front end of the IS-641 codec, which was introduced in Section 1.2.1.

Distributions of the LSF parameters are presented in Figure 2.7. The LSF parameters vary between 89 Hz and 3775 Hz. It can be noticed that the variation of LSF parameters is most evident in the middle frequencies. This results from basic properties of the speech; the formats are essential for recognizing phonemes and typically the formats in the middle

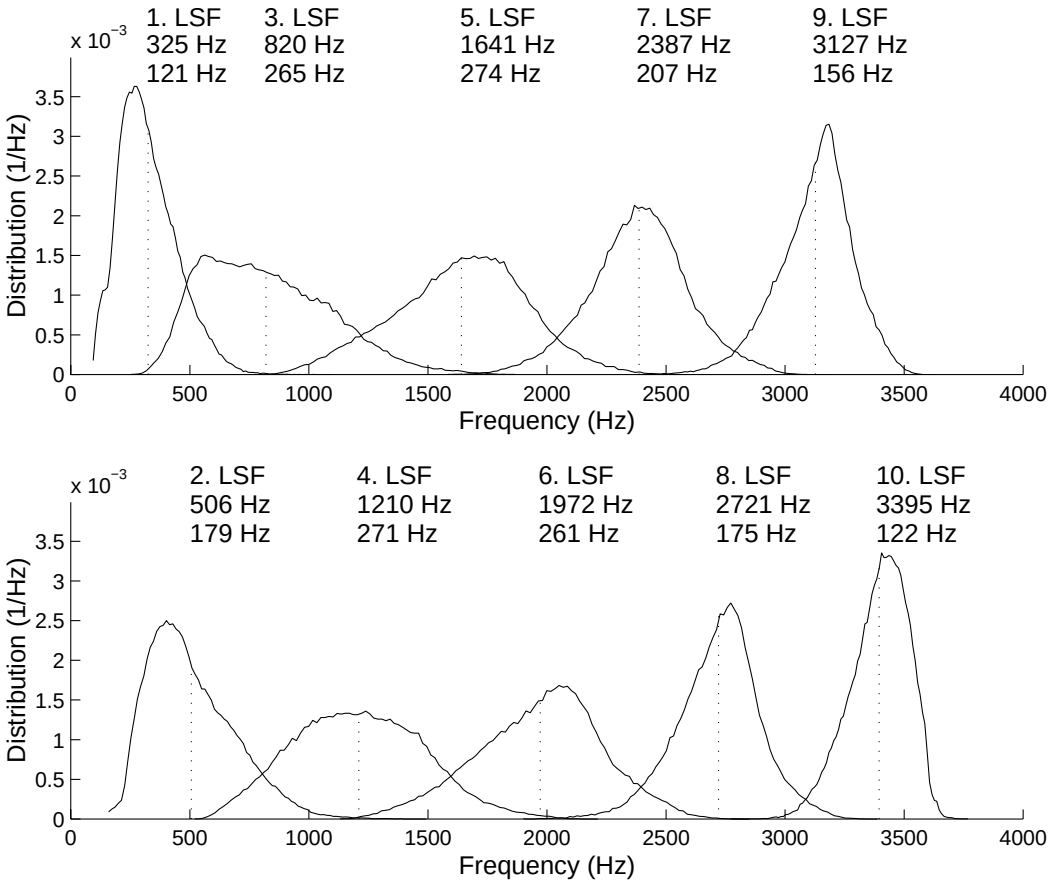

Figure 2.7. Distributions of the LSF parameters in the training database.
The average value and the standard deviation are presented below the number of LSF.

frequency area vary considerably depending on the pronounced phoneme, so these variations affect to the middle part of magnitude response and to middle LSF parameters. Also average values are not evenly divided but are emphasized towards lower frequencies. The magnitude response calculated from the average values of the LSF parameters is presented in Figure 2.8. The magnitude response increases from 0 Hz to 400 Hz and then decreases approximately 6 dB per octave until reaching 4000 Hz.

A correlation matrix of the LSF parameters calculated from the training database is shown in Table 2.1. It indicates a strong correlation between neighboring LSF parameters, near a diagonal of the correlation matrix. The correlations are strongest between the lowest LSF parameters. Also the correlations of the LSF parameters in adjacent frames are strong, as can be seen in Table 2.2. The correlations of the LSF parameters inside a frame are important for efficient vector quantization. The correlation between adjacent frames can be utilized by using predictive quantization as will be discussed in Chapter 3.

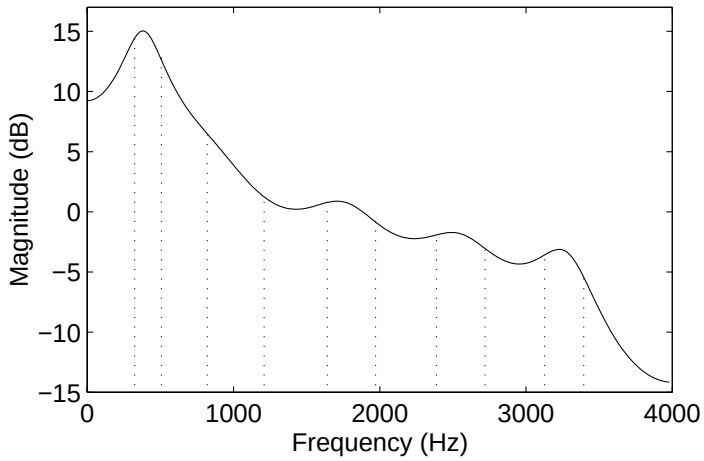

Figure 2.8. Magnitude response calculated from the average values of the LSF parameters in the training database. The average values are displayed with dotted lines.

Table 2.1. The correlation of the LSF parameters.

LSF	1	2	3	4	5	6	7	8	9	10
1	1.00	0.83	0.62	0.46	0.24	0.14	0.19	0.14	0.07	–0.17
2	0.83	1.00	0.80	0.55	0.32	0.19	0.23	0.22	0.13	–0.09
3	0.62	0.80	1.00	0.77	0.43	0.35	0.33	0.33	0.26	0.04
4	0.46	0.55	0.77	1.00	0.64	0.46	0.42	0.32	0.30	0.07
5	0.24	0.32	0.43	0.64	1.00	0.78	0.53	0.40	0.22	–0.02
6	0.14	0.19	0.35	0.46	0.78	1.00	0.69	0.47	0.35	0.08
7	0.19	0.23	0.33	0.42	0.53	0.69	1.00	0.70	0.46	0.28
8	0.14	0.22	0.33	0.32	0.40	0.47	0.70	1.00	0.54	0.24
9	0.07	0.13	0.26	0.30	0.22	0.35	0.46	0.54	1.00	0.60
10	–0.17	–0.09	0.04	0.07	–0.02	0.08	0.28	0.24	0.60	1.00

Table 2.2. The correlation of the LSF parameters in adjacent frames.
The LSF parameters of the current frame are in the first column and past frame are in the first row.

LSF	1	2	3	4	5	6	7	8	9	10
1	0.76	0.69	0.59	0.46	0.25	0.17	0.22	0.18	0.11	–0.11
2	0.62	0.76	0.70	0.51	0.30	0.19	0.23	0.24	0.15	–0.03
3	0.47	0.63	0.80	0.65	0.37	0.31	0.30	0.32	0.25	0.08
4	0.36	0.43	0.63	0.82	0.55	0.40	0.37	0.29	0.28	0.10
5	0.18	0.24	0.36	0.56	0.86	0.70	0.47	0.36	0.20	0.01
6	0.09	0.13	0.27	0.39	0.69	0.85	0.61	0.42	0.31	0.09
7	0.13	0.16	0.26	0.35	0.46	0.60	0.82	0.59	0.40	0.25
8	0.09	0.16	0.25	0.25	0.34	0.40	0.58	0.79	0.44	0.21
9	0.03	0.08	0.19	0.24	0.18	0.30	0.39	0.45	0.77	0.49
10	–0.19	–0.11	–0.01	0.02	–0.04	0.05	0.21	0.18	0.46	0.74

2.5 Objective distortion measures

The human auditory system is the ultimate evaluator of the quality of a speech coder and the performance in preserving intelligibility and naturalness. However, extensive subjective tests are often too time consuming and expensive and a testing setup that is independent of a particular coding or analysis-synthesis system is difficult to define. Thus, several objective measures for LP filter quantization are presented in literature (see for example [23]). The objective distortion measures used in this thesis to evaluate the quality of quantization in speech coding, are presented in the following subsections.

2.5.1 Spectral distortion

The spectral distortion (SD) measure is most commonly used in speech coding for measuring the LP filter quantization performance. The SD is defined in dB units as follows [24]:

$$SD_i = \sqrt{\frac{1}{f_h - f_l} \int_{f_l}^{f_h} \left[10\log_{10}(P_i(f)) - 10\log_{10}(\hat{P}_i(f)) \right]^2 df} \tag{2.8}$$

where f_h and f_l define the upper and lower frequency (Hertz) limits for integration. Symbols $P_i(f)$ and $\hat{P}_i(f)$ denote the LP power spectra of the ith frame given by

$$P_i(f) = 1 \Big/ \left| A_i(e^{j2\pi f/F_s}) \right|^2, \tag{2.9}$$

$$\hat{P}_i(f) = 1 \Big/ \left| \hat{A}_i(e^{j2\pi f/F_s}) \right|^2, \tag{2.10}$$

where F_s is the sampling frequency, and $A_i(z)$ and $\hat{A}_i(z)$ are the original unquantized and the quantized LPC polynomials, respectively. Originally f_h is equal to 0 Hz and f_l corresponds to half of the sampling frequency, but in practice spectral distortion is often calculated over a limited bandwidth. In this thesis f_h and f_l are equal to 80 Hz and 3600 Hz, respectively. This choice corresponds to the passband filtering of the speech signal.

In literature the average SD for all frames in an evaluation database is used together with some measures on outliers, i.e., fraction of frames where SD exceeds a given threshold. According to Paliwal and Atal [9], *transparent quality* is attained when the average SD is about 1 dB and the percentage of outlier frames having SD between 2 and 4 dB is less than 2 %. Moreover, no frames must have SD greater than 4 dB. By transparent quantization of LPC parameters, it is meant here that the quantization does not introduce any audible distortion in the coded speech; i.e. coded speech with and without quantization of the LPC parameters are indistinguishable through listening.

In this thesis the average SD is denoted by SD_{ave}, the percentage of outliers having SD over 2 dB by SD_{2dB} and the percentage of outliers having SD over 4 dB by SD_{4dB}. Since the SD_{4dB} is typically extremely low (nearly zero) for feasible quantizers, it is rarely used in this thesis. Usually if SD_{2dB} is over the limit, then also the limit of SD_{4dB} is exceeded.

2.5.2 Weighted Euclidean distance

When the codebooks for the quantizers are being designed, the complexity of calculating the SD measure makes it impractical for use. Since the LSF representation has a substantial relationship with the shape of the spectral envelope it is natural to use an Euclidean distance measures. The weighted Euclidean distance measure is defined as

$$d_{ED}(\mathbf{x},\hat{\mathbf{x}}) = (\mathbf{x} - \hat{\mathbf{x}})^T \mathbf{W} (\mathbf{x} - \hat{\mathbf{x}}), \tag{2.11}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are the original and quantized LSF vectors, respectively, and $\mathbf{W}$ is a symmetric and positive definite weighting matrix that may depend on $\mathbf{x}$.

The weighting matrix $\mathbf{W}$ is often designed to emphasize perceptually significant vector elements. Thus the weighting allows for the quantization of the LSF parameters in the formant regions better than those in the non-formant regions. Various methods for calculating $\mathbf{W}$ have been proposed by several authors [9, 15, 25, 26, 27], but the performances of measures are essentially similar [28]. In this thesis, we employ the weighting matrix $\mathbf{W}$ whose diagonal elements w_i^2 are obtained as in the IS-641 codec:

$$w_i = \begin{cases} 3.347 - \dfrac{1.547}{450} d_i, & \text{for } d_i < 450 \\ 1.8 - \dfrac{0.8}{1050}(d_i - 450), & \text{for } d_i \geq 450 \end{cases}. \tag{2.12}$$

Here $d_i = x_{i+1} - x_{i-1}$ with $x_0 = 0$ Hz and $x_{11} = 4000$ Hz. The symbol x_i denotes the ith line spectral frequency in Hertz. However, the weighting is not used in the training phase of the quantizer in order to quarantee the stability of the quantized LP filters [9], i.e., the weighting matrix is set to an identity matrix $\mathbf{W} = \mathbf{I}$. In this thesis the d_{ED} value is used to denote the mean of several values of Equation (2.11). The value is presented in unit rad^2/s^2.

2.5.3 Average quantization error

An average quantization error in Hertz is given by

$$d_{AQ} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(x_i - \hat{x}_i)^2}\,, \tag{2.13}$$

where k is the dimension of coded vectors. The Average Quantization Error measure is used in Chapters 3 and 4 to compare split quantizers ability to code corresponding LSF parameters. As the Weighted Euclidean value, the Average Quantization Error value is used as a mean of several values of Equation (2.13).

2.5.4 Segmental signal-to-noise ratio

Since the SD, d_{ED} and d_{AQ} measures are used in the frequency domain, the most commonly used time domain measure, the signal-to-noise ratio, is presented below. The signal-to-noise ratio measures the relative strength of signal power with respect to noise power. However, the simple signal-to-noise ratio weights all time domain errors in the signal according to their energies, neglecting the fact that speech energy is time-varying. Thus it does not give an accurate estimate of speech quality [29]. A segmental signal-to-noise ratio measure overcomes this problem. It is the geometric mean of signal-to-noise measures conducted on a frame-by-frame basis. The segmental signal-to-noise ratio measure, denoted here by segSNR, in decibels, over K speech segments is defined as

$$\text{segSNR} = \frac{1}{K}\sum_{k=0}^{K-1} 10\log_{10}\left(\frac{\sum_{n=1}^{L} s^2(n+Lk)}{\sum_{n=1}^{L}(s(n+Lk)-\hat{s}(n+Lk))^2}\right), \tag{2.14}$$

where $\hat{s}(n)$ are $s(n)$ the coded and original speech samples, respectively, and each segment k has a length of L samples. The value 160 is used for L in this thesis to correspond to the frame length of the IS-641 speech codec.

3 Moving Average Vector Quantization

Over the years, research activity on the quantization of LPC parameters has been very extensive. The principal objective in quantization is to avoid any audible distortion in coded speech while coding the LPC parameters at as low a bit rate as possible. If this objective is met, *transparent quality* or *transparent quantization* is said to be achieved.

Since the 70s a number of quantization techniques have been reported in literature. At that time first scalar quantizers were used. These techniques quantize individual parameters separately, using either uniform or nonuniform quantizers. Since the late 70s research of vector quantization (VQ) has grown rapidly following the increasing computational power that made VQ a practical alternative [30, 31, 32]. VQ can be seen as an extension of scalar quantization to a multidimensional space. VQ considers the entire set of parameters as an entity and quantizes the entire vector at once as a scalar quantizes one parameter at a time independently. Shannon [33, 34] has shown that for a given bit rate, coding longer blocks of information will always attain better performance in terms of lower distortion.

A huge amount of papers has been published on VQ. VQ has shown to be more efficient than scalar quantization of spectral parameters [16, 32], since coded parameters correlate. The correlation of LSF parameters was shown and discussed in detailed in Sections 2.3 and 2.4. A more recent tendency is to utilize interframe prediction of spectral parameters [35, 36]. Here coding efficiency can be improved by exploiting the temporal redundancy of spectral envelopes. The present vector quantizers, usually coding LSF parameters, require an accuracy of about 20–30 bits per frame to achieve transparent quality, while the scalar quantizers require 32–50 bits per frame.

However, VQ is not without problems. The size of codebook of quantizer consisting of the possible quantized vectors increases exponentially with the number of bits. This affects to computational complexity and memory requirements. With current technology, the upper limit of a bits for practical use is 10–14, i.e. the maximum number of codevectors in codebook is about 1000–16000. Since an optimal unconstrained VQ is not feasible, sub-optimal constrained quantizers are used. Split vector quantization (SVQ) and multi-stage VQ are the most commonly used techniques, in which the overall quantization is divided into two or more smaller tasks.

Another problem in VQ is codebook optimization. Since an analytical solution of codebook optimization does not exist, iterative algorithms have to be used. In this thesis two- and three-split vector quantizers and their codebook optimizations are considered. Moving average predictors are used in these quantizers, and their structures and optimization are also studied.

In Section 3.1 the vector quantizer and its common design tool, *Generalized Lloyd algorithm* (GLA) is defined. Since the correlations of LSF parameters are strong between adjacent frames, predictive quantizers are efficient. A moving average split vector quantizer (MA-SVQ) is presented in Section 3.2. The section also defines predictor structures used in this thesis; diagonal matrix, full matrix and inter-split predictor. In the last section, optimal prediction coefficients are solved and iterative algorithms for training MA-SVQ are presented.

3.1 Vector quantization

Since the spectral parameters are usually nonuniformally distributed, like LSF in Figure 2.7 in Section 2.4, nonuniformal quantizers based on codebooks are used. The vector quantizer, or encoder, maps a k-dimensional input vector $\mathbf{x}$ to a codevector $\mathbf{y}_i$ satisfying

$$d(\mathbf{x},\mathbf{y}_i) \le d(\mathbf{x},\mathbf{y}_j),\ j \ne i, \tag{3.1}$$

where $d(\,)$ is a suitable distortion measure. $\mathbf{y}_i$ and $\mathbf{y}_j$ are codevectors of a codebook $\mathbf{C} = \{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_N\}$, where N is a number of the codevectors in the codebook. Since the codebook $\mathbf{C}$ is fixed, only the index i is transmitted to decoder and an appropriate output of the decoder is $\hat{\mathbf{x}} = \mathbf{y}_i$.

The design of codebooks is usually accomplished by an iterative training algorithm. A set of representative vectors of the source is compiled for training, and the codebook is optimized using a suitable distortion measure. An optimal VQ has to satisfy two necessary conditions [37, 38, 39]; the Nearest Neighbor Condition and the Centroid Condition.

The Nearest Neighbor Condition says that, given a decoder and its finite set of output codevectors $\mathbf{C}$, the encoder's optimal partition cells $\{R_i\}$ satisfy

$$R_i \subset \left\{ \mathbf{x} \;\middle|\; d(\mathbf{x}, \mathbf{y}_i) \le d(\mathbf{x}, \mathbf{y}_j)\ ;\ \forall\, j \right\}. \tag{3.2}$$

This means that for a given $\mathbf{x}$ the encoding region R_i with codevector $\mathbf{y}_i$ is the optimal choice and no other codevector $\mathbf{y}_j$ can give less distortion.

The Centroid Condition says that, given an encoder's partition $P = \{ R_i \mid i = 1, \ldots, N \}$, the optimal codevectors $\mathbf{y}_i$ in $\mathbf{C}$ are the centroids in each partition cell R_i:

$$\mathbf{y}_i \;=\; \mathrm{cent}(R_i) \;=\; \arg\min_{\mathbf{y}} E\left[d(\mathbf{x}, \mathbf{y}) \;\middle|\; \mathbf{x} \in R_i \right]. \tag{3.3}$$

3.1.1 Generalized Lloyd algorithm

The *Generalized Lloyd algorithm* (GLA) is the most widely used algorithm to design a VQ codebook, see [37]. GLA, also known as the LBG algorithm [40], is an iterative clustering algorithm for optimal VQ design based on training vectors. The algorithm satisfies both the Nearest Neighbor Condition and the Centroid Condition. The algorithm is shown in Figure 3.1.

1. Initialize the codebook **C**.
2. For each vector **x** in training database:
   (i) Find the codevector $\mathbf{y}_i$ that gives the smallest distortion according to (3.2).
   (ii) Update the estimates of the expectations in (3.3).
3. Abortion test. Test if the iterations have converged or apply some other abortion test to find out if the optimization is finished. If not, continue from 2.
4. Termination.

Figure 3.1. GLA algorithm.

A problem with iterative algorithms is that they can get trapped in a local minimum. Thus the initial values of codebook are a crucial step for efficient VQ mapping. Several methods have been proposed in literature [37, 40, 41] to avoid this problem. In this thesis a simple solution is used to avoid the problem: multiple initial values have been tried and the best codebook according to the used distortion measure is chosen.

The size of the training set and the number of GLA iterations are critical factors during the training process. The set should be large enough to closely approximate the statistical characteristics of the vector sequence. If a small amount of data is used for training the quantizer it learns the set too detailed and suffers a lack of efficiency with data outside the training set. A reasonable thumb rule is that the ratio of the training set vectors to the

number of codebook vectors should be above 50 [32]. The number of GLA iterations should be limited so that the codebook would not be excessively trained. If the codebook is overly trained, it will again perform poorly with data outside the training set. Consequently testing the performance of the vector quantizer is done on a separate validation data that were not used during training.

Another problem with training of a quantizer is the performance outside the training set. In speech applications, quantizers usually perform satisfactorily when used on the same type of data as that in the training set. However, the performance may be reduced significantly when used with other types of speech material. In this thesis, the training data consist on flat and modified IRS filtered speech data while the validation data consist on IRS filtered data (see Section 2.2). This mismatch of filtering is used to simulate the real life environment, where the used data is not similar to that used in the training.

The most commonly used distortion measure for training a codebook is the Euclidean distance measure of LSF parameters, i.e., a squared error between the original and quantized values. Although the SD measure, Equation (2.8), would be a better objective measure the high computational complexity makes it impractical for codebook design and codebook search at encoder. Paliwal and Atal [9] observed problems with stability when the weighted Euclidean distance measure was used in the GLA and since they used an unweighted measure. Following Paliwal and Atal, Euclidean distance, Equation (2.11), where weighting matrix is replaced with an identity matrix $\mathbf{W} = \mathbf{I}$, is used in the training process in this thesis.

However, the weighted Euclidean distance is used in the encoding process. While there is now a mismatch between the encoding and codebook training, Hagen [42] has reported that they compared the results obtained with unweighted and weighed measures used in the design of codebooks and did not observe any differences for an evaluation set.

3.2 Moving average split vector quantization

There is a high degree of similarity among the spectral envelopes of neighboring speech frames, especially within the vowels. In Section 2.4 it was shown that the LSF parameters have high correlation between adjacent frames (see Table 2.2 and Figures 2.2–2.4). Consequently, highly efficient quantization can be realized by utilizing the interframe correlations. This is especially useful in the case of low bit rate and short frame coder.

A number of predictive quantizers have been proposed [36, 37, 43, 44, 45, 46, 47, 48], which typically utilize moving average (MA), autoregressive (AR) or non-linear predictors. Unfortunately, prediction that is based on recursive reconstructions of the decoder can suffer from the propagation of channel errors over numerous frames. In MA predictive

coding, the error propagation is limited to number of frames given by the order of the MA predictor. Contrary, in AR predictive coding the error propagation is infinite, which makes it undesirable for real life speech coders. Linear prediction, like MA and AR prediction, is optimal when the data is stationary and Gaussian distributed. Since LSF parameters are not Gaussian distributed, non-linear predictors can provide an accurate model to the spectral parameters [43, 44, 46, 47, 48]. However, the non-linear predictors suffer from complexity, stability problems, and suitable training algorithms, and thus are not considered in this thesis.

As earlier mentioned, a single VQ system, producing transparent quality of speech spectrum, requires a large codebook which leads to intractable complexity. Therefore the overall quantization of an input vector is divided into smaller tasks. Thus two- and three-split quantizers are explored in this thesis. For two-split quantizer, the 10-dimensional LSF vector is partitioned into subvectors of dimensions 4 and 6. For the three-split vector quantizer, the LSF vector is split into subvectors of dimensions 3, 3 and 4, respectively.

3.2.1 General moving average vector quantizer

The general moving average split vector quantizer (MA-SVQ) is defined as

$$\hat{\mathbf{x}}(t) = \mathbf{B}_0\mathbf{u}(t) + \mathbf{B}_1\mathbf{u}(t-1) + \cdots + \mathbf{B}_{n_B}\mathbf{u}(t-n_B), \tag{3.4}$$

where $\hat{\mathbf{x}}(t)$ is a quantized vector, n_B is an order of the predictor, $\mathbf{B}_i$ is a prediction coefficient matrix of an order i and $\mathbf{u}(t-i)$ is a chosen code vector from a codebook $\mathbf{C}$ at time instant $t-i$. Here the vectors and the matrix are partitioned to s splits,

$$\hat{\mathbf{x}}(t) = \begin{bmatrix} \hat{\mathbf{x}}_1(t) \\ \vdots \\ \hat{\mathbf{x}}_s(t) \end{bmatrix}, \quad \mathbf{B}_i = \begin{bmatrix} \mathbf{B}_{i11} & \cdots & \mathbf{B}_{i1s} \\ \vdots & & \vdots \\ \mathbf{B}_{is1} & \cdots & \mathbf{B}_{iss} \end{bmatrix} \text{ and } \mathbf{u}(t-i) = \begin{bmatrix} \mathbf{u}_1(t-i) \\ \vdots \\ \mathbf{u}_s(t-i) \end{bmatrix}, \tag{3.5}$$

where $\mathbf{u}_k(t-i)$ and $\hat{\mathbf{x}}_k(t)$ are m_k-vectors and the submatrix $\mathbf{B}_{ijk}$ of the coefficient matrix $\mathbf{B}_i$ is an m_j by m_k matrix. The codebook $\mathbf{C}$ comprises separate codebooks $\{\mathbf{C}_1, ..., \mathbf{C}_s\}$. A codebook $\mathbf{C}_k = \{\mathbf{u}_k^{(1)}, \mathbf{u}_k^{(2)}, \ldots, \mathbf{u}_k^{(N_k)}\}$, where $\mathbf{u}_k^{(i)}$ is a i-th codevector of the split k and N_k is a number of codevectors in the split k. In this thesis the number of the codevectors in codebook $\mathbf{C}_k$ is presented in bits, thus a bit rate of the sub-quantizer of split k defined $N_{\text{bit}k} = \log_2(N_k)$. The overall bit rate of the quantizer, N_{bit}, is a sum of the bit rates of the sub-quantizers.

There are several possible structures for the prediction coefficient matrices $\mathbf{B}_0$, ..., $\mathbf{B}_{n_B}$. The choice of prediction matrix structure depends on computational complexity, efficiency,

memory usage and training algorithm of predictor. Three different predictor structures that are used in this thesis, are presented in the following sub-sections.

3.2.2 Diagonal matrix predictor

Firstly, in the simplest predictor structure $\mathbf{B}_0$ is an identity matrix and the matrices $\mathbf{B}_1$, …, $\mathbf{B}_{n_B}$ are diagonal. Thus the quantizer can be divided to s independent sub-quantizers, where s is a number of splits. Equation (3.4) can be rewritten for every sub-quantizer k, where $1 \leq k \leq s$,

$$\hat{\mathbf{x}}_k(t) = \mathbf{u}_k(t) + \mathbf{B}_{1kk}\mathbf{u}_k(t-1) + \cdots + \mathbf{B}_{n_B kk}\mathbf{u}_k(t-n_B). \tag{3.6}$$

Here the prediction matrix $\mathbf{B}_i$, where $1 \leq i \leq n_B$, is divided to diagonal submatrices $\mathbf{B}_{ikk}$

$$\mathbf{B}_{ikk} = \begin{bmatrix} b_{ik1} & 0 & \cdots & 0 \\ 0 & b_{ik2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & b_{ikm_k} \end{bmatrix} \tag{3.7}$$

Since there are no connections between splits, each split can be independently quantized. In fact, there are no connections of neighboring parameters either. Thus the memory and calculation requirements for predictor are slight, only $10n_B$ prediction parameters have to be stored and $10n_B$ multiplications and additions have to be made for the prediction of one 10-dimensional LSF vector. The MA-SVQ using diagonal matrix predictor is denoted as n_B-DP-s-SVQ. The integer s is optional and used only when quantizers having separate numbers of splits are discussed. For example, a three-split quantizer using first order diagonal matrix predictor would be denoted 1-DP-3-SVQ or 1-DP-SVQ depending on context.

3.2.3 Full matrix predictor

Secondly, a more complex predictor structure can be defined from Equation (3.6), when full submatrices $\mathbf{B}_{ikk}$, where $1 \leq i \leq n_B$, are used instead of diagonal matrices. Thus every sub-quantizer k can be defined as

$$\hat{\mathbf{x}}_k(t) = \mathbf{u}_k(t) + \mathbf{B}_{1kk}\mathbf{u}_k(t-1) + \cdots + \mathbf{B}_{n_B kk}\mathbf{u}_k(t-n_B), \tag{3.7}$$

where $1 \leq k \leq s$, and the full matrix predictor is defined as

$$\mathbf{B}_{ikk} = \begin{bmatrix} b_{ik11} & \cdots & b_{ik1m_k} \\ \vdots & & \vdots \\ b_{ikm_k1} & \cdots & b_{ikm_km_k} \end{bmatrix}. \tag{3.8}$$

The memory and calculation requirements of the predictor increases compared to the diagonal matrix predictor. In the case of the three-split quantizer of 10 parameters, $34n_B$ prediction parameters have to be stored and $34n_B$ multiplications and additions have to be made. In the case of the two-split quantizers the corresponding numbers are $52n_B$ and $52n_B$. The MA-SVQ using the full matrix predictor is denoted as n_B-FP-s-SVQ, where s is an optional parameter.

3.2.4 Inter-split predictor

Thirdly, the inter-split predictor is the most elaborate structure of these predictors. The prediction of vector is based on intra- and inter-split parameters. Thus the prediction of $\mathbf{x}_k(t)$ is not only based on the quantized parameters of same split, as in diagonal and full matrix predictor, but also to the quantized parameters of the other splits. Actually, all known parameters of codevectors $\mathbf{u}(t-i)$, where $0 \le i \le n_B$, are utilized to make prediction of subvector $\mathbf{x}_k(t)$, where $1 \le k \le s$. Therefore chosen codevector parameters of the same frame can also be utilized for making prediction.

The inter-split vector quantizer is defined in Equation (3.4). The prediction matrices $\mathbf{B}_1$, …, $\mathbf{B}_{n_B}$ are full. However matrix $\mathbf{B}_0$ has a special structure, containing full, diagonal, and zero-element submatrices. The order of these submatrices defines the quantization order of the splits. In this thesis, ascending and declining quantization orders of the splits are used.

In the first case, the quantization order of splits is ascending. Thus the vector of first split $\mathbf{x}_1(t)$ is quantized with prediction from quantized inter-frame vectors $\mathbf{u}(t-i)$, where $1 \le i \le n_B$. The following recursion is used for other splits; $\mathbf{x}_k(t)$, where k goes from 2 to s, is quantized with prediction from quantized intra-frame subvectors $\mathbf{u}_j(t)$, where $1 \le j < k$, and quantized inter-frame vectors $\mathbf{u}(t-i)$. In this case the prediction matrix $\mathbf{B}_0$ is defined as

$$\mathbf{B}_0 = \begin{bmatrix} \mathbf{B}_{011} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{B}_{021} & \mathbf{B}_{022} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{0} \\ \mathbf{B}_{0s1} & \cdots & \mathbf{B}_{0s(s-1)} & \mathbf{B}_{0ss} \end{bmatrix}, \tag{3.9}$$

where $\mathbf{B}_{0kk}$ are unit matrices and $\mathbf{B}_{0jk}$, when $j > k$, are full matrices. The upper half of the matrix $\mathbf{B}_0$ has zero elements.

In the second case, the quantization order of splits is declining, starting from $\mathbf{x}_s(t)$ using recursion where k diminishes from s to 1, respectively. The prediction matrix $\mathbf{B}_0$ is defined as

$$\mathbf{B}_0 = \begin{bmatrix} \mathbf{B}_{011} & \mathbf{B}_{012} & \cdots & \mathbf{B}_{01s} \\ \mathbf{0} & \mathbf{B}_{022} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{B}_{0(s-1)s} \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{B}_{0ss} \end{bmatrix} \tag{3.10}$$

where $\mathbf{B}_{0kk}$ are unit matrices and $\mathbf{B}_{0jk}$, when $j < k$, are full matrices. The lower half of the matrix $\mathbf{B}_0$ has zero elements.

The MA-SVQ using the inter-split predictor is denoted as n_B-IS$_{12\ldots s}$-SVQ in the first, ascending case and n_B-IS$_{s(s-1)\ldots1}$-SVQ in the latter, declining case. For example three-split IS-SVQs using first order predictor are denoted 1-IS$_{123}$-SVQ and 1-IS$_{321}$-SVQ. In the case where quantization order of splits is not fixed quantizers are denoted either n_B-IS-s-SVQ or n_B-IS-SVQ.

Since matrices $\mathbf{B}_1$, …, $\mathbf{B}_{n_B}$ are full and matrix $\mathbf{B}_0$ is nearly half filled, the memory and the calculation requirements are demanding. Memory usage of the prediction parameters and the number of multiplications and additions are nearly $100(n_B + 1)$, which is ten times more than the demand of the diagonal predictor structure. Beside the training of the inter-split predictor is complex and slow. However, the rapid development of integrated circuit technology will make it a practical choice. Finally, the inter-split predictor is the most efficient structure of these, at least theoretically, since it utilizes all correlation information available.

3.3 Optimization of MA-SVQ

There exist several algorithms for optimizing the codevectors and prediction coefficients for MA-SVQ. Typically these optimization algorithms are either Lloyd-based or gradient descent iterative algorithms. In this thesis an Lloyd-based algorithm is used. Although an analytical solution of the optimal codevectors does not exist, the prediction parameters can be solved analytically. Since the Lloyd algorithm can be slightly modified to take account the explicitly solved prediction parameters.

3.3.1 Prediction parameter estimation

The predictor coefficients can be explicitly solved, assuming that codevectors $\mathbf{u}(t)$, where $t = 1, 2, \ldots, M$ and M is a number of training vectors, are known and fixed. In other words, original vectors $\mathbf{x}(t)$ have to be encoded to get codevectors $\mathbf{u}(t)$, then the optimal predictor coefficients can be solved. Firstly, all parameters of coefficient matrices $\mathbf{B}_0$, …, $\mathbf{B}_{n_B}$ are stacked columnwise into a parameter vector $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \begin{bmatrix} \operatorname{col} \mathbf{B}_0 \\ \vdots \\ \operatorname{col} \mathbf{B}_{n_B} \end{bmatrix}, \tag{3.11}$$

and the codevectors $\mathbf{u}(t)$, …, $\mathbf{u}(t - n_B)$ are stacked columnwise into a vector $\boldsymbol{\varphi}(t)$,

$$\boldsymbol{\varphi}(t) = \begin{bmatrix} \mathbf{u}(t) \\ \vdots \\ \mathbf{u}(t - n_B) \end{bmatrix}. \tag{3.12}$$

Thus the quantized vector $\hat{\mathbf{x}}(t)$ in Equation (3.7) can be rewritten

$$\hat{\mathbf{x}}(t) = \Phi(t)\boldsymbol{\theta}, \tag{3.13}$$

where

$$\Phi(t) = \boldsymbol{\varphi}^{\mathrm{T}}(t) \otimes \mathbf{I}, \tag{3.14}$$

and operator $\otimes$ is the Kronecker product. Here $\boldsymbol{\theta}$ is a $100(n_B + 1)$-vector, $\boldsymbol{\varphi}(t)$ is a $10n_B$-vector and $\Phi(t)$ is a $10n_B$ by $100(n_B + 1)$ matrix.

Secondly, a selector vector $\mathbf{q}$ is defined. The selector $\mathbf{q}$ is a $100(n_B + 1)$-vector having binary elements. An i-th element q_i of $\mathbf{q}$ has value 1, if corresponding element θ_i of the parameter vector $\boldsymbol{\theta}$ is free and value 0, if θ_i is fixed. The element of $\boldsymbol{\theta}$ is free, if it can be optimized. Respectively it is fixed, if it has either value zero, as for example with off-diagonal elements of $\mathbf{B}_k$ of diagonal matrix predictor, or value one, as diagonal elements of $\mathbf{B}_0$. Consequently the selector $\mathbf{q}$ has mostly zero elements in the case of diagonal matrix predictor and respectively mostly unit elements in the case of inter-split predictor.

Thirdly, a free parameter, defined by selector $\mathbf{q}$, of the vector $\boldsymbol{\theta}$ can be picked up to a vector $\boldsymbol{\theta}'$. Thus the vector $\boldsymbol{\theta}'$ has size $\|\mathbf{q}\|$, which is between [0, $100(n_B + 1)$]. Similarly, a

matrix $\Phi'(t)$ can be defined from the matrix $\Phi(t)$, collecting only columnvectors, where the corresponding selector element q_i of $\mathbf{q}$ has value 1. $\Phi'(t)$ is a $10n_B$ by $\|\mathbf{q}\|$ matrix. Equation (3.7) has now the form

$$\hat{\mathbf{x}}(t) = \Phi'(t)\boldsymbol{\theta}'. \tag{3.15}$$

The optimal predictor coefficients can be solved from Equation (3.15). The objective function for minimization is given by

$$J(\boldsymbol{\theta}', \mathbf{C}) = (1/M)\sum_{t=1}^{M} \left\| \mathbf{x}(t) - \Phi'(t, \mathbf{C})\boldsymbol{\theta}' \right\|^2, \tag{3.16}$$

where the Euclidean measure is used. Minimizing Equation (3.16) with respect to the parameter vector $\boldsymbol{\theta}'$ gives an estimate

$$\boldsymbol{\theta}' = \mathbf{R}^{-1}\mathbf{r}, \tag{3.14}$$

where

$$\mathbf{R}^{-1} = \Big(\sum_{t=1}^{M} \Phi'(t)^{\mathrm{T}}\Phi'(t)\Big)^{-1},$$

$$\mathbf{r} = \sum_{t=1}^{M} \Phi'(t)^{\mathrm{T}}\mathbf{x}(t). \tag{3.15}$$

The parameters of coefficient matrices $\mathbf{B}_0, \ldots, \mathbf{B}_{n_B}$ can now be picked from the parameter vector $\boldsymbol{\theta}'$ using the selector $\mathbf{q}$ and Equation (3.11).

3.3.2 Training algorithm of MA-SVQ

MA-SVQ is trained using optimal prediction parameter estimation and Generalized Lloyd algorithm (GLA) in turn. First the GLA algorithm is used to train several codebooks for every split (or memoryless sub-quantizers). Then the codebooks $\mathbf{C}_k$, $1 \le k \le s$, having smallest quantization error, are collected to an initial codebook $\mathbf{C} = \{\mathbf{C}_1, \ldots, \mathbf{C}_s\}$. The prediction parameters are solved using Equations (3.14) and (3.15). In the next step, the GLA algorithm has been modified slightly to include the predictor. Therefore a prediction error vector is defined

$$\begin{aligned}\mathbf{e}(t) &= \mathbf{x}(t) - \hat{\mathbf{x}}(t) + \mathbf{u}(t), \\ &= \mathbf{x}(t) - \mathbf{B}_0\mathbf{u}(t) - \mathbf{B}_1\mathbf{u}(t-1) - \cdots - \mathbf{B}_{n_B}\mathbf{u}(t-n_B) + \mathbf{u}(t). \end{aligned} \quad (3.16)$$

Instead of $\mathbf{x}$, the prediction error $\mathbf{e}$ is used in the GLA algorithm in Figure 3.1. Every split is encoded using Equation (3.2) and the codebooks $\mathbf{C}_k$ are updated with Equation (3.3). When a inter-split predictor is being trained, the error vector $\mathbf{e}$ is updated after a codebook update in each split. The prediction parameter estimation and codebook updating are performed by turns, until the algorithm has converged. The training algorithm is shown in Figure 3.2.

1. Initialize the codebook **C** using GLA. Set prediction matrices $\mathbf{B}_i$ to zero.
2. Solve optimal prediction parameters for $\mathbf{B}_i$.
3. Decode and calculate prediction error vector **e**.
4. Update the codebook **C,** i.e., the estimates of the expectations in Equation (3.3), where **e** is used instead of **x**.
5. Encode splits, i.e., find the codevectors **u** that yields the smallest distortion according to Equation (3.2), where prediction error **e** is used instead of **x**.
6. Abortion test. Test if the iterations have converged or apply some other abortion test to find out if the optimization is finished. If not, continue from 2.
7. Termination.

Figure 3.2. Training algorithm for MA-SVQ.

The algorithm is iterative and sub-optimal; the parameters are optimized by turns, expecting that the other parameters are fixed. Nevertheless the algorithm converges, as it will be seen in Chapter 4. However, the convergence of the algorithm is much slower than in a case of the memoryless vector quantizer. In this thesis the algorithm was allowed to iterate, until no further convergence was noticed. During the iteration the best codebooks and prediction parameters were stored. After the termination of the algorithm the best codebooks and prediction parameters are chosen. However, more elaborate abortion test could easily be implemented.

4 Objective Results

This chapter presents the objective test results for the moving average split vector quantizers (MA-SVQs). The reference for this study is the LSF quantizer of the IS-641 codec. The quantizer of IS-641 is three-split vector quantizer using the first order diagonal predictor (1-DP-SVQ). The bit allocation of splits is 8, 9 and 9 bits, resulting in 26 bits per frame. Therefore this chapter evaluates three-split LSF quantizers of bit rates from 14 to 26 bits per frame. Furthermore the two-split quantizers of lower bit rates are included in the study. Diagonal matrix, full matrix and inter-split predictors are used in these quantizers to enhance their performance.

Most commonly used measure for evaluating the quantizers is the spectral distortion which was introduced in Section 2.5. Since lower spectral distortion values can be attained using the weighting matrix of from Equation (2.12) in quantization, weighting is used to guide in the encoding process of every quantizer. The spectral distortion values are presented mostly only for the validation database. In addition, the spectral distortion values for the training database and the average quantization error values for both databases are presented in some example cases. Altogether the results of nearly two hundred split vector quantizers are shown in this chapter. Since the amount of data is enormous, only the most important results are included.

The objectives of this chapter are, firstly, to compare the performance of different predictor structures and secondly, to define the overall objective quality of quantizers. Two main references to assess the quantizers are the limits for transparent quality of LPC parameters and the quality of the quantizer of IS-641. According to Paliwal and Atal [9], transparent quality is achieved if SD_{ave} is about 1 dB and SD_{2dB} is less than 2 %.

Details of the training phase of quantizers are discussed in the next section. The section presents iteration curves of the training phase and evaluates the required computation work. Section 4.2 shows the results of reference quantizers, IS-641 and memoryless. Following Sections 4.3 and 4.4 present the objective results of the predictive three and two-split quantizers, respectively. The last section summarizes the results. The section also suggests three alternative quantizers to IS-641 and compares performance of these quantizer.

4.1 Training of MA-SVQ

The Lloyd-based training algorithm, presented in Section 3.5, is used to train predictive quantizers examined in this chapter. Quantizers with diagonal and full matrix predictors can be independently trained for each split. This makes the training phase relatively straightforward; let the algorithm iterate, and after its termination, pick the best sub-quantizer. Thus the choice of sub-quantizers for the desired bit rate can be done afterwards and several sub-quantizer combinations can be tested to achieve the optimal solution.

However, a quantizer using an inter-split predictor has to be trained for every split simultaneously. Because of this the numbers of codevectors in each split have to be decided in advance. The full matrix quantizer is a suitable for an initial value for the training. The codebooks and prediction coefficients of sub-quantizers can be trained in turns. The memory requirement of the algorithm with large training data is huge. Since split codebooks are not independent, careful planning and manual guidance of the training algorithm are required to accelerate the iteration process. Especially two-split IS-SVQs get easily trapped to a local minimum. Generally the training of IS-SVQs training process is far more complex and slow compared to the training of other predictor structures. For these reasons limited number of the quantizers using inter-split predictors are trained in this thesis.

4.1.1 Training time

The time needed to train a quantizer depends on the number of iterations, computational burden of the algorithm and the available computation power. In Figure 4.1, typical iteration curves of the training phase of predictive quantizer are shown. As it can be noticed, the algorithm converges first rapidly after starting and then saturates slowly. Occasionally, after some oscillations the iteration curve diverges. Because of this the training algorithm stores always the quantizer which gives the smallest quantization error. The predictor affects the convergence curve so that more complicated structures of predictor usually converge slower than simple ones, as it can be seen from the figure. Typically the number of iterations needed varies from 20 to several hundreds.

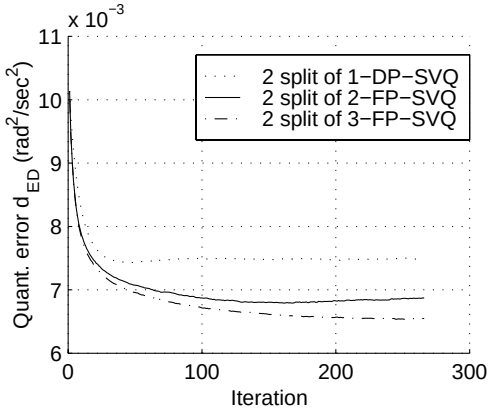

Figure 4.1. Iteration curves for the training phase of the second split of two-split 18-bit quantizers. Nine bits is allocated for both splits.

Table 4.1. Estimated floating point calculations of one iteration of the training algorithm in training database. The bit rate of quantizer per frame is N_{bit} and n_B is a order of predictor.

	Number of additions	*Number of multiplications*
2-split	$(3.0 \cdot 2^{N_{\text{bit}}/2} + 50 \cdot {n_B}^2) \cdot 10^6$	$(3.0 \cdot 2^{N_{\text{bit}}/2} + 50 \cdot {n_B}^2) \cdot 10^6$
3-split	$(3.7 \cdot 2^{N_{\text{bit}}/3} + 30 \cdot {n_B}^2) \cdot 10^6$	$(3.7 \cdot 2^{N_{\text{bit}}/3} + 30 \cdot {n_B}^2) \cdot 10^6$

Computation work of one iteration is estimated in Table 4.1. The estimates are based on the training algorithm in the training database. The estimates are mostly functions of bit rate of quantizer. Nevertheless the order of predictor has minor effect, especially with lower bit rates. The number of floating point additions and multiplications is practically the same. For 18-bit quantizers, with full matrix predictor of order one, the number of additions and multiplications in one iteration is $1.6 \cdot 10^9$ for a two-split quantizer and $0.31 \cdot 10^9$ for a three-split quantizer. When the predictor is full and the order is three, the corresponding numbers are $2.1 \cdot 10^9$ and $0.58 \cdot 10^9$. As it can be seen from Table 4.1, the training of the two-split quantizers requires considerably more computation. The encoding of the training data impacts the most to the computation, especially at higher bit rates. Therefore if the training time should be minimized, the most efficient solution is to enhance the encoder. One solution might be some kind of accelerated search from codevectors. The examples for such algorithms can be found in [37].

The final thing that affects the training time is the computation power. Some of the quantizers were trained with either a computer equipped with a 133 MHz Pentium processor or a computer equipped with a 450 Pentium II processor. An estimated computation power of the 133 MHz Pentium was six millions double precision floating point additions and multiplications per second. The comparison of these two computers shows that the computational power has grown more than quadrupled. However, in practice it was learned that the training time depends mostly on the bit rate. For example, the 133 MHz Pentium could perform one iteration loop of 17-bit quantizer of three splits, with

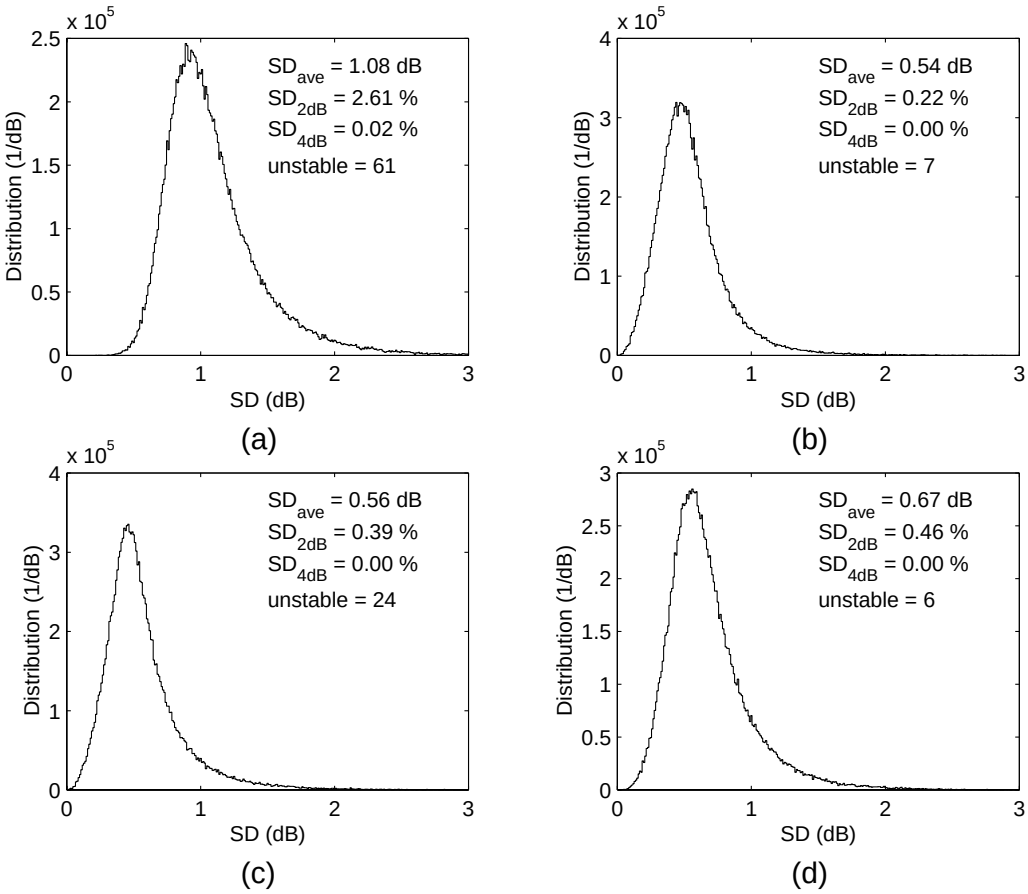

Figure 4.2. Spectral distortion distributions for 3-FP-SVQ of 23 bits when (a) all splits, (b) only first, (c) only second, and (d) only third split are quantized. The distributions have been computed from the training database. The bit allocation of splits is 7, 8, and 8, respectively.

predictor of order one, in one minute. However, it took two minutes with the 450 MHz Pentium II to perform one loop of similar quantizer having 26 bits.

4.1.2 Optimal bit allocation

Finally before presenting the results, optimal bit allocations of sub-quantizers of splits are discussed. The LSF parameters in middle frequencies vary more than LSF parameters of high and low frequencies (see Section 2.4), thus more bits are needed to quantize them accurately. In addition, the LSF parameters of low frequencies are nearer to each other than the higher ones, so small quantization errors have higher effect on the spectrum. Thus bit allocation also emphasizes quantizers of the first split. Nevertheless any rule for the optimal bit allocation does not exist and therefore it has to be found experimentally.

As an example of a bit allocation, typical spectral distortion distributions of a 23-bit quantizer are shown in a Figure 4.2. The first split has seven bits and other ones eight bits each. The effect of quantization error of individual split can be seen in graphs (b), (c) and

(d), where the spectral distortion is measured, while only one split is quantized and other splits remain unquantized. The third split has 4 LSF parameters to encode, consequently this affects the most in the spectral distortion values.

4.2 Objective results of reference quantizers

The 26-bit LSF quantizer adopted from the IS-641 codec and memoryless quantizers of bit rates 14–29 bits are used as references to the quantizers developed in this thesis. The LSF quantizer of IS-641 is denoted by $DP_{IS\text{-}641}$-SVQ, since it uses the first order diagonal matrix predictor (1-DP-SVQ). The $DP_{IS\text{-}641}$-SVQ has three splits whose bit allocation is 8, 9, and 9, producing 26 bits in total. The original $DP_{IS\text{-}641}$-SVQ has been trained only with modified IRS filtered data unlike the quantizers developed in this thesis. The memoryless or not predictive quantizers are trained with the training database using the generalized Lloyd algorithm presented in Figure 3.1. The sub-quantizers of each split are trained separately and their optimal combinations are found experimentally.

Figures 4.3 and 4.4 present spectral distortion of the quantizers for the validation and training databases. The two-split memoryless quantizers (2-SVQs) obtain better results than the three-split memoryless quantizers (3-SVQs). The performance gain of the two-split quantizers is approximately one bit in the validation database and two bits in the training database. The limits of transparent quality are met with a 29-bit 3-SVQ. The results of the memoryless quantizers are nearly two bits lower in the validation database than in the training database.

The $DP_{IS\text{-}641}$-SVQ fulfills the limits of transparency quality in the validation database. However as Figure 4.1 show, $DP_{IS\text{-}641}$-SVQ suffers strongly when used with unsuitable data. Table 4.2 clarifies the matter. The week performance of the quantizer results mostly from the unfiltered (flat) data in the last row of Table. This weakens the overall result of the training database considerably, since the results of only modified IRS filter part of the training database fulfill the transparency limits. The performance decreases mostly in the first split as it can be seen from d_{AQ1} values.

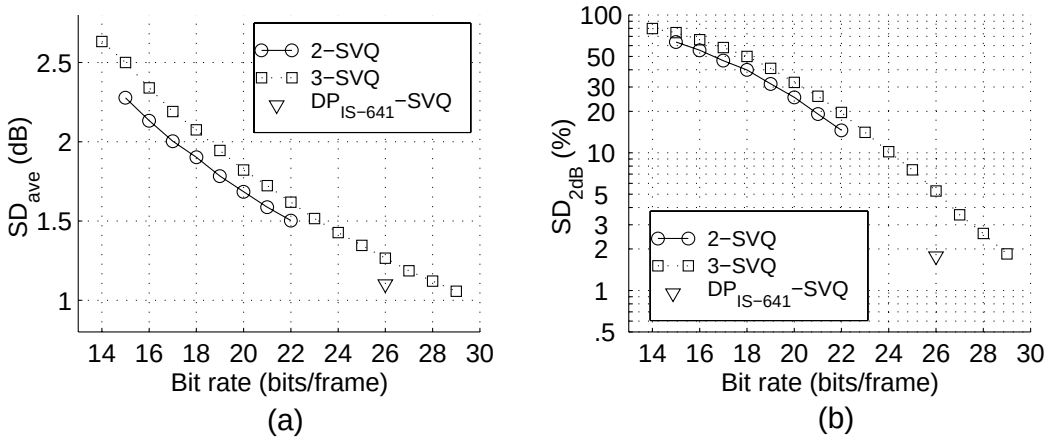

Figure 4.3. SD values of the memoryless SVQs and the quantizer of IS-641 for the validation database.

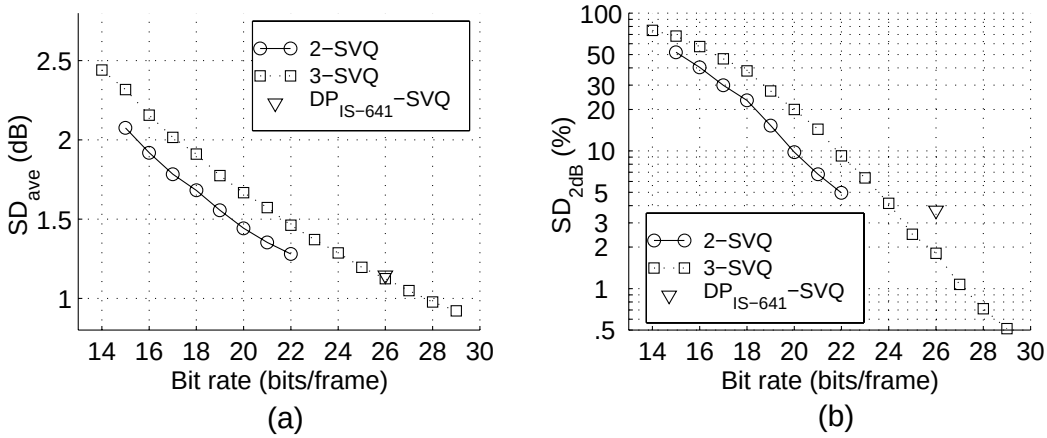

Figure 4.4. SD values of the memoryless SVQs and the quantizer of IS-641 for the training database.

Table 4.2. Objective measures for the quantizer of IS-641 in the validation, the training and both the modified IRS (Tr. modIRS) and the flat (Tr. flat) filtered parts of the training database. The number of unstable filters is denoted by Unst.

Database	SD_{ave} (dB)	SD_{2dB} (%)	SD_{4db} (%)	Unst.	d_{AQ} (Hz)	d_{AQ1} (Hz)	d_{AQ2} (Hz)	d_{AQ3} (Hz)
Validation	1.10	1.8	0.0	5	25.4	20.3	22.5	28.0
Training	1.15	3.7	0.0	36	27.4	23.0	24.8	28.5
Tr. ModIRS	1.05	1.3	0.0	17	24.8	19.8	23.2	26.4
Tr. Flat	1.24	6.1	0.0	19	30.0	26.2	26.4	30.5

4.3 Objective results of three-split quantizers

This section is divided into four parts. In the first part objective results of quantizers using diagonal predictors (DP-SVQs) are studied. Spectral distortion is presented for both the

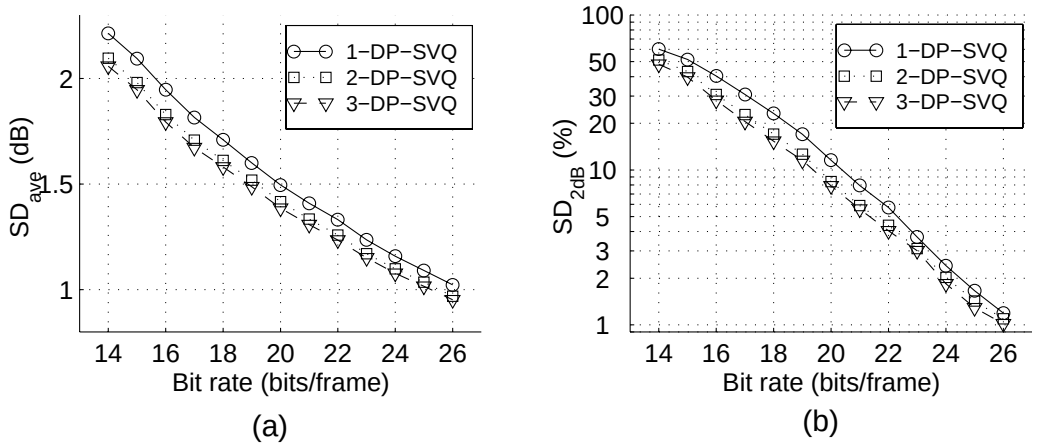

Figure 4.5. SD values of the three-split quantizers using diagonal matrix predictors for the validation database.

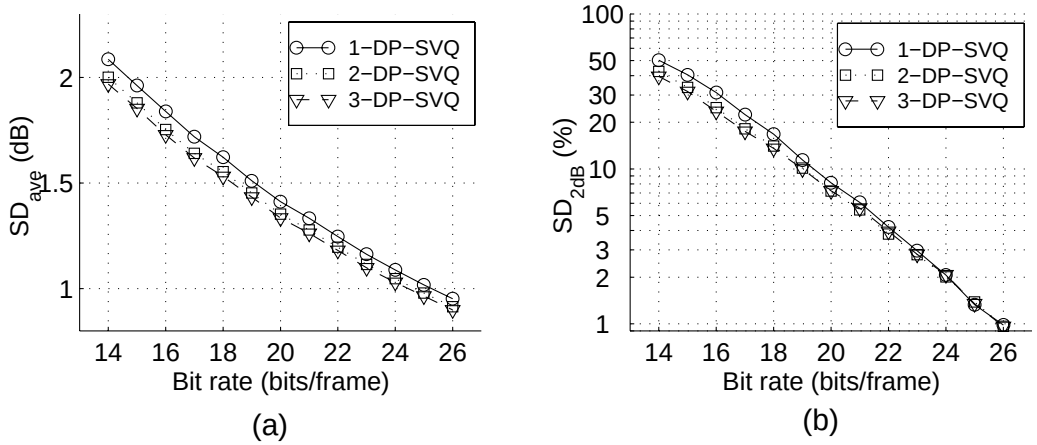

Figure 4.6. SD values of the three-split quantizers using diagonal matrix predictors for the training database

.

training and the validation databases. In addition detailed results of 1-DP-SVQs are presented for the validation database. In the subsequent sections, spectral distortion of quantizers using full matrix (FP-SVQs) and inter-split (IS-SVQs) predictors, respectively, are presented for the validation database. In the last section, the objective measures are compared with each other and to the results of the reference quantizers.

4.3.1 Three-split quantizers using diagonal matrix predictor

The SD_{ave} and the SD_{2dB} values of DP-SVQs for the validation database are presented in Figure 4.5. Several observations can be made based on these results. Firstly, the SD_{ave} values behave rather linearly as a function of bit rate. Secondly, 3-DP-SVQs give an average of 7 % smaller spectral distortion than 1-DP-SVQs. The corresponding performance gain is approximately one bit. Thirdly, 2-DP-SVQs outperform 1-DP-SVQs, on average by 5 % in terms of spectral distortion. Finally, when the bit rate diminishes by one bit, spectral distortion increases roughly 0.1 dB.

Figure 4.6 shows corresponding SD values for the training database. The shapes of the curves are similar as with the validation database. However the SD_{ave} values of lower bit rates are nearly 0.1 dB lower than the SD_{ave} values of the validation data and with higher bit rates the difference converges to 0.05 dB. Here 1-DP-SVQs are outperformed by 3-DP-SVQs, on average by 5.5 % and 2-DP-SVQs, on average by 4 %.

The detailed objective measures of 1-DP-SVQs in the validation database are shown in Table 4.3. Table presents the bit allocations of the quantizers, SD values, number of frames with unstable filter and average quantization error d_{AQ} values. The second column indicates that in the optimal bit allocation the sub-quantizer of first split gets by at least one bit lower than the sub-quantizers of other splits. The SD values show that only the 25-bit and the 26-bit quantizers are fairly near the limits of transparent quality. Although the SD_{ave} values are slightly above 1.0 dB limit, it has to keep in mind that the training and validation databases have different statistical properties because of unequal prefilterings (see Section 2.2). Thus the SD_{ave} values of 25- and 26-bit quantizers in the training database are 1.02 and 0.95, respectively. The number of unstable frames and SD_{4dB} values decrease rapidly when bit rate increases. With bit rates above 20 bits neither of number of unstable frames nor SD_{4dB} value are problematic. The operation of the sub-quantizers can be studied from the average quantization error values in the last four columns. The average error of all LSF parameters d_{AQ} diminishes linearly as the bit rate increases. The average quantization errors of splits; d_{AQ1}, d_{AQ2} and d_{AQ3}, have nearly similar values on average. However with higher bit rates the average error of first split d_{AQ1} is smaller than d_{AQ2} and d_{AQ3} values.

Table 4.3. Objective measures of the three-split 1-DP-SVQs in the validation database. The bit allocation of the quantizers are shown in Alloc.-column. The number of unstable filters is denoted by Unst.

Bits	Alloc.	SD_{ave} (dB)	SD_{2dB} (%)	SD_{4db} (%)	Unst.	d_{AQ} (Hz)	d_{AQ1} (Hz)	d_{AQ2} (Hz)	d_{AQ3} (Hz)
14	4,5,5	2.21	60.3	0.9	87	52.9	47.7	51.7	51.5
15	4,6,5	2.09	51.7	0.6	59	49.7	47.7	41.3	51.5
16	5,6,5	1.95	40.4	0.4	52	46.6	37.0	41.3	51.5
17	5,6,6	1.82	30.7	0.2	47	42.7	37.0	41.3	43.0
18	5,6,7	1.71	23.2	0.2	45	39.8	37.0	41.3	35.9
19	5,7,7	1.60	17.0	0.1	31	37.1	37.0	32.8	35.9
20	6,7,7	1.50	11.5	0.1	19	34.8	29.6	32.8	35.9
21	6,7,8	1.41	7.9	0.1	18	32.4	29.6	32.8	30.3
22	6,8,8	1.33	5.7	0.1	13	30.4	29.6	26.4	30.3
23	7,8,8	1.24	3.7	0.1	6	28.5	23.4	26.4	30.3
24	7,8,9	1.16	2.4	0.0	8	26.5	23.4	26.4	25.6
25	7,9,9	1.09	1.7	0.0	5	24.8	23.4	21.2	25.6
26	8,9,9	1.02	1.2	0.0	3	23.5	18.8	21.2	25.6

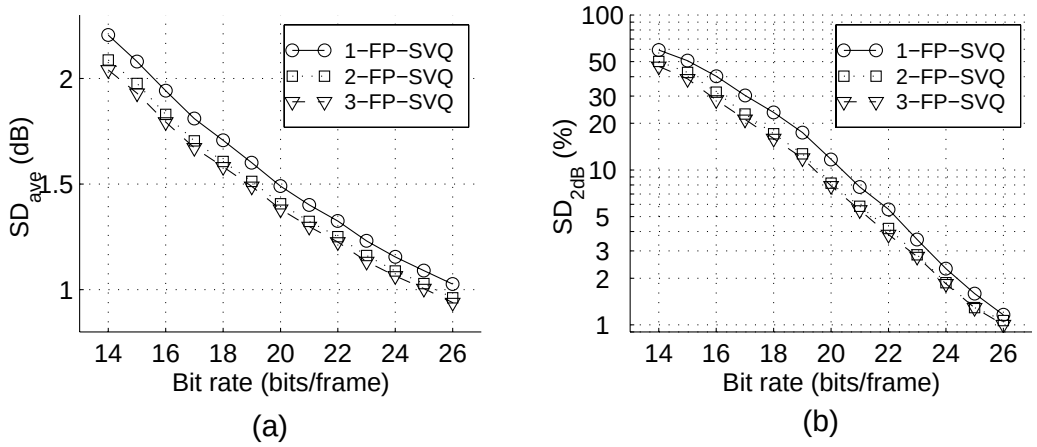

Figure 4.7. SD values of the three-split quantizers using full matrix predictors for the validation database.

4.3.2 Three-split quantizers using full matrix predictor

The previous part presented the SD values for both the training and validation database. It was noticed that the SD_{ave} values of the validation database are an average of 0.05–0.1 dB higher than corresponding values of the training database. In addition the SD and the average quantization error values correspond to the validation database of 1-DP-SVQs. For these reasons, only the SD results of the validation database are shown in this part.

Figure 4.7 shows the SD values of FP-SVQs. The results are essentially similar than the results of DP-SVQs. Here 3-FP-SVQs outperform 1-FP-SVQs an average of 7.5 % in the SD_{ave} values and 2-FP-SVQs are 5.5 % better than 1-FP-SVQs. As with the case of DP-SVQs, 3-FP-SVQs have approximately one bit performance gain compared to 1-FP-SVQs. Furthermore in higher bit rates 2-FP-SVQs have similar SD_{ave} values than 1-FP-SVQs using one bit more. For example the 26-bit 1-FP-SVQ has SD_{ave} equal to 1.03 dB and SD_{2dB} equal to 1.2 % and the 25-bit 2-FP-SVQ has SD_{ave} equal to 1.03 dB and SD_{2dB} equal to 1.3 %.

4.3.3 Three-split quantizers using inter-split predictor

Since the training process of IS-SVQs is complex only 14, 17, 20 and 23–26-bit IS-SVQs are evaluated. The SD values of these quantizers in the validation database are presented in Figure 4.8. Shapes of the SD value curves are similar to the corresponding curves of DP-SVQs and FP-SVQs. However IS-SVQs have an average of 5 % smaller SD_{ave} values than DP-SVQs and FP-SVQs. Therefore the inter-split predictor structure gains more than 0.5 bit saving to the bit rate compared to other predictors. The quantization order of splits does not seem to have effect to the results, the performance of IS_{123}-SVQs and IS_{321}-SVQs are rather similar. The 2-IS-SVQs have nearly the same performance than 1-IS-SVQs using one bit more.

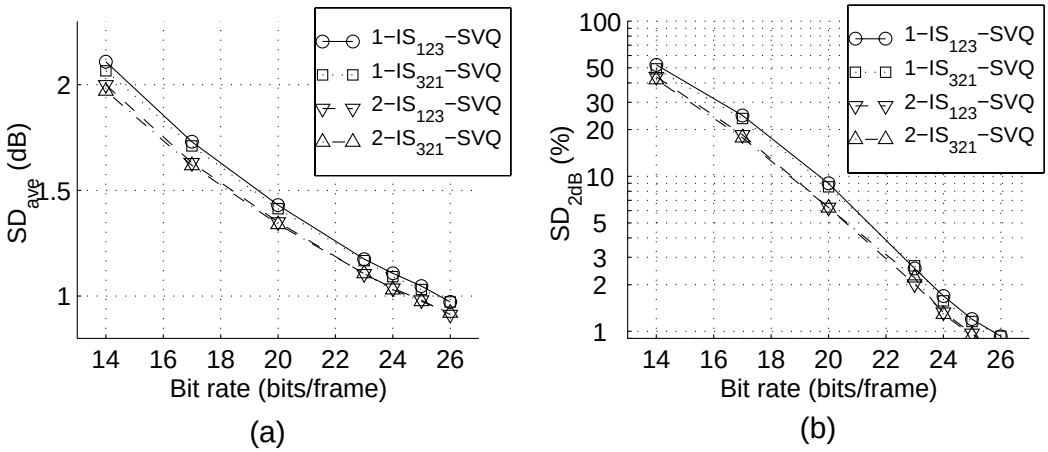

Figure 4.8. SD values of the three-split quantizers using inter-split predictors for the validation database.

4.3.4 Comparison of three-split quantizers

The SD values of the reference and the previously presented quantizers are presented in Figure 4.9 for the validation database and in Figure 4.10 for the training database. The IS-SVQs achieve the best results. The limits of transparent quality and the performance of $DP_{IS\text{-}641}$-SVQ can be met with the 23-bit 2-IS-SVQ and the 24-bit 1-IS-SVQ. The 25-bit 1-DP-SVQ outperforms $DP_{IS\text{-}641}$-SVQ. The memoryless quantizer needs 29 bits to achieve both the limits of transparent quality and the performance of $DP_{IS\text{-}641}$-SVQ.

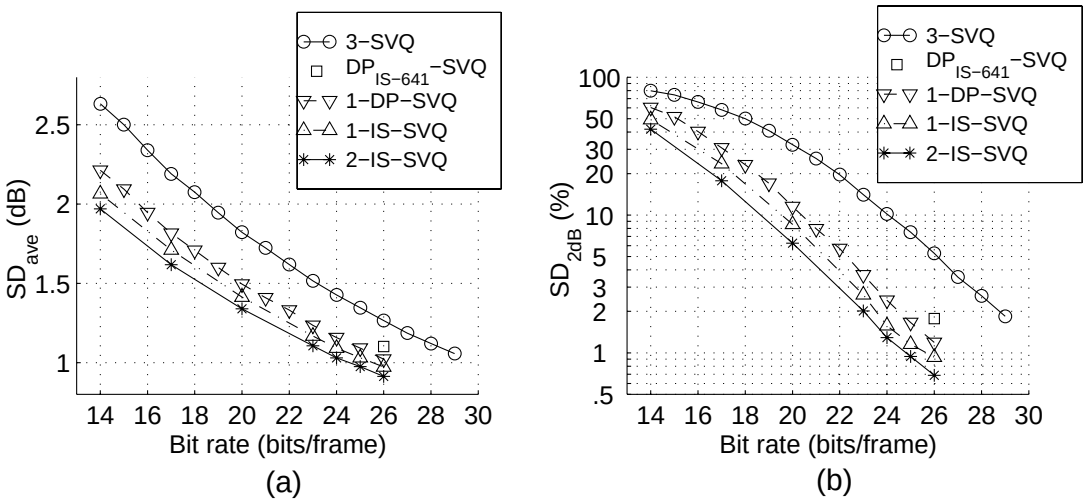

Figure 4.9. Comparison of SD values of the three-split quantizers for the validation database.

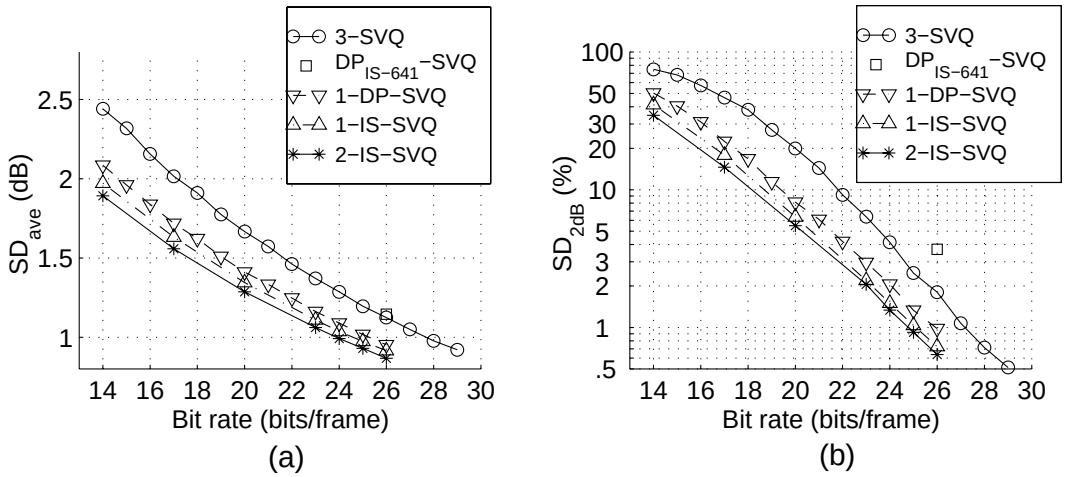

Figure 4.10. Comparison of SD values of the three-split quantizers for the training database.

The previous conclusions motivate to study relative bit rates of the quantizers. The relative bit rate estimate of two quantizers is calculated comparing differences of the SD_{ave} values and the bit rates of these quantizers. Table 4.4 presents the relative bit rate estimates of the three-split quantizers compared to 1-DP-SVQs. Firstly, the differences of the relative bit rate estimates are larger in the validation than in the training database. This results from the fact that the SD values of simpler quantizer (as memoryless quantizer) weaken more than the SD values of sophisticated quantizer (as IS-SVQ) when using unsuitably filtered data. Thus, it seems that the complicated predictive quantizers tolerate the unsuitable data better than the simple predictive or the memoryless quantizers. Secondly, the relative bit rate estimates of DP-SVQs and FP-SVQs are nearly equivalent, which shows that their performances are equal. However, IS-SVQs outperform DP-SVQs and FP-SVQs, on average by 0.7–0.8 bits. Thirdly, the table shows that the maximum bit save, nearly 1.5 bits, can be achieved using 2-IS-SVQs instead of 1-DP-SVQs.

Table 4.4. The relative bit rate estimates of the 14–29 bits three-split quantizers compared to 1-DP-SVQs. The estimates are calculated from the validation and the training databases.

Quantizer	*Rel. bit rate (valid.)*	*Rel. bit rate (train.)*
3-SVQ	-3.18	-2.47
1-DP-SVQ	0.00	0.00
1-FP-SVQ	0.04	0.06
1-IS-SVQ	0.95	0.76
2-DP-SVQ	0.87	0.65
2-FP-SVQ	0.95	0.76
2-IS-SVQ	1.79	1.46
3-DP-SVQ	1.16	0.87
3-FP-SVQ	1.26	1.02

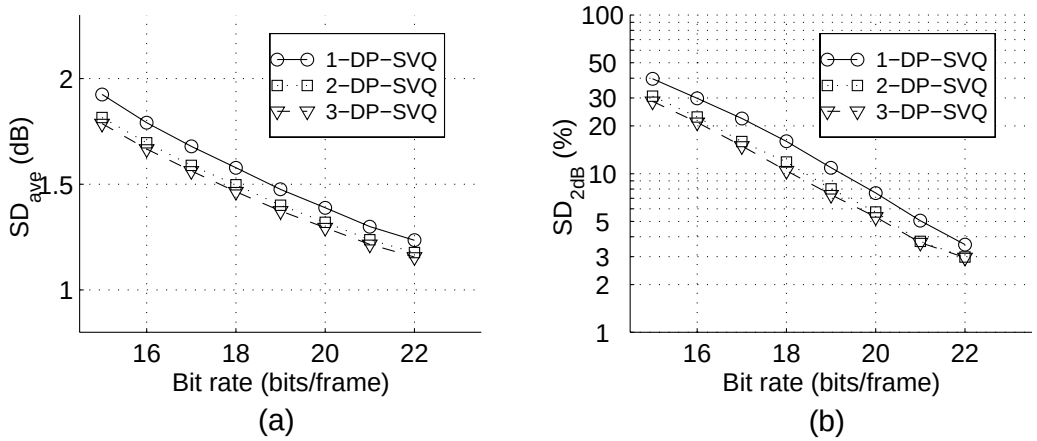

Figure 4.11. SD values of the two-split quantizers using diagonal matrix predictors for the validation database.

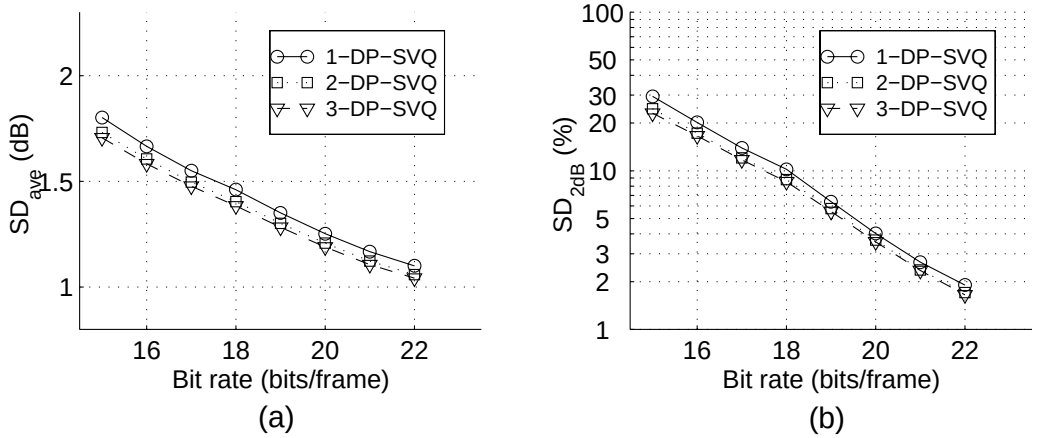

Figure 4.12. SD values of the two-split quantizers using diagonal matrix predictors for the training database.

4.4 Objective results of two-split quantizers

Following the structure of the previous section, the results of two-split quantizers are presented in four parts. The quantizers of 15–22 bits are evaluated. In the first part, the SD values of DP-SVQs are presented for both the training and the validation databases. In the second and the third parts, the SD values of FP-SVQs and IS-SVQs, respectively, are presented for the validation database. In the last part, the objective measures are compared with each other.

4.4.1 Two-split quantizers using diagonal matrix predictor

The SD_{ave} and the SD_{2dB} values of DP-SVQs are presented in Figure 4.11 for the validation database and in Figure 4.12 for the training database. None of the quantizers fulfill the

requirements of transparent quality. The best quantizer, 22-bit 3-DP-SVQ, obtains SD_{ave} equal to 1.16 dB and SD_{2dB} equal to 3.0 % in the validation database. However the results of the training database are promising. The 22-bit quantizers reach the transparency limits in the training database, since 1-DP-SVQ has SD_{ave} equal to 1.10 dB and 2-DP-SVQ and 3-DP-SVQ have SD_{ave} around 1.05 dB and SD_{2dB} values of these quantizers are all below 2 % limit.

The 3-DP-SVQs have an average of 7 % better SD_{ave} values than 1-DP-SVQs in the validation database. So 3-DP-SVQs have the same performance as 1-DP-SVQs using one bit more. The 2-DP-SVQs outperform 1-DP-SVQs, on average by 5 % in the SD_{ave} values in the validation database. Furthermore the SD_{ave} values of the validation database are an average of 0.1–0.15 dB higher than corresponding values of the training database.

Table 4.5 shows the detailed objective measures of 1-DP-SVQs in the validation database. The second column indicates that in optimal bit allocation the sub-quantizer of first split gets by at several bits fewer than the sub-quantizers of second split. Since the second split has six LSFs it requires plenty of bits for quantization. Although it is known that the last LSF parameters 9 and 10 tolerate can be quantized loosely than the other parameters [9] the training algorithm does not use this possibility. So to enhance the performance of the two-split quantizer the use of weighting matrix in Equation (2.12) also in the training phase could enhance the performance. However this was not done in this thesis cause of the reasons described in Section 2.5. The another interesting observation is that the SD_{ave} values saturate towards the 22-bit quantizer. It appears that a 22-bit quantizer could obtain better SD values if bit allocation would be 9 and 13. However this is not reasonable since the complexity of quantizer would not be tolerable. Finally with the bit rates above 20 bits, neither of the number of unstable frames nor the SD_{4dB} value are problematic. Nevertheless the SD_{ave} and the SD_{2dB} values are slightly too high for transparent quality in the two-split DP-SVQs.

Table 4.5. Objective measures of the two-split 1-DP-SVQs in the validation database. The bit allocation of the quantizers are shown in Alloc.-column. The number of unstable filters is denoted by Unst. .

Bits	Alloc.	SD_{ave} (dB)	SD_{2dB} (%)	SD_{4dB} (%)	Unst.	d_{AQ} (Hz)	d_{AQ1} (Hz)	d_{AQ2} (Hz)
15	7,8	1.93	39.6	0.3	51	46.0	36.3	50.1
16	7,9	1.79	29.9	0.2	33	42.3	36.3	44.5
17	7,10	1.68	22.3	0.1	21	39.3	36.3	39.7
18	8,10	1.58	16.0	0.1	20	37.0	30.3	39.7
19	8,11	1.48	10.9	0.0	12	34.3	30.3	35.4
20	8,12	1.39	7.6	0.0	9	32.0	30.3	31.7
21	9,12	1.30	5.0	0.0	8	30.2	25.6	31.7
22	10,12	1.24	3.6	0.0	5	28.8	21.8	31.7

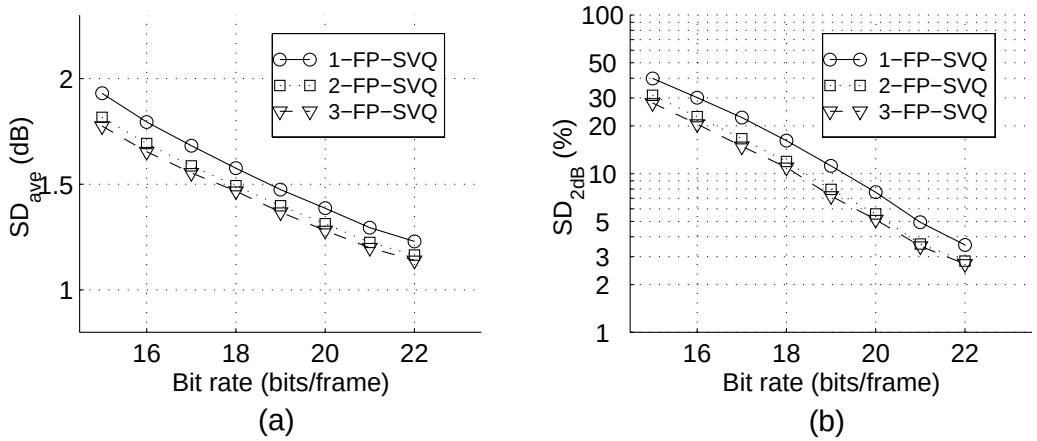

Figure 4.13. SD values of the two-split quantizers using full matrix predictors for the validation database.

4.4.2 Two-split quantizers using full matrix predictor

Figure 4.13 shows the SD values of FP-SVQs in the validation database. The results are similar to the results of DP-SVQs. The SD_{ave} values of FP-SVQs differ less than 1 % from corresponding values of DP-SVQs. Here 3-FP-SVQs outperform 1-FP-SVQs, on average by 7.5 % in the SD_{ave} values. Both 2-FP-SVQs and 3-FP-SVQs have the same performance as 1-FP-SVQs using one bit more. The 22-bit 3-FP-SVQ has SD_{ave} equal to 1.14 dB and SD_{2dB} equal to 2.7 % in the validation database and corresponding values in the training database are 1.04 dB and 1.6 %.

4.4.3 Two-split quantizers using inter-split predictor

Two-split IS-SVQs of 16, 18, 20 and 22 bits are evaluated. Figure 4.14 shows the SD values of IS-SVQs in the validation database. The IS-SVQs have an average of 2.5 % smaller SD_{ave} values than DP-SVQs and FP-SVQs. Thus, in the two-split case, IS-SVQs are not as efficient compared to the other predictor structures, as in the three-split case, where the corresponding value is 5 %. However IS-SVQs gain more than 0.5 bit saving to the bit rate compared to other quantizer structures. In each case, the quantization order of splits is declining (IS_{21}-SVQ). Finally, 2-IS-SVQs have nearly the same performance than 1-IS-SVQs using one bit more.

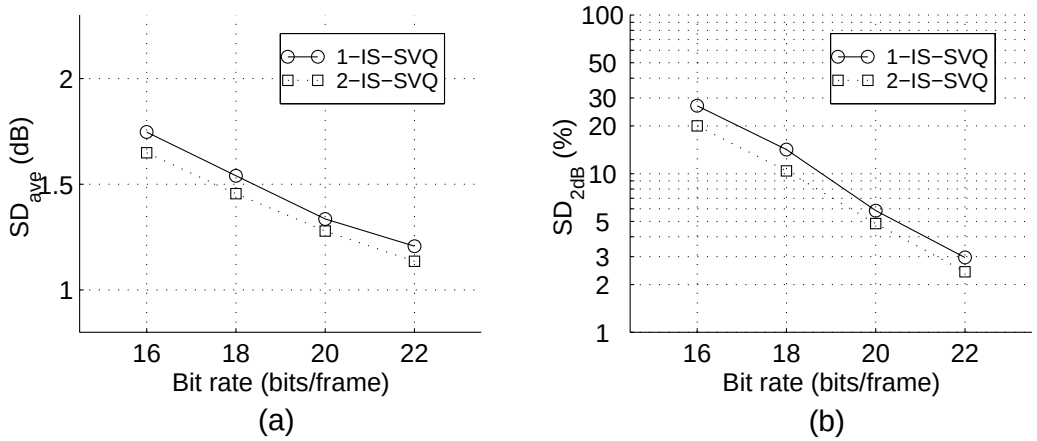

Figure 4.14. SD values of the two-split quantizers with inter-split predictors for the validation database.

4.4.4 Comparison of two-split quantizers

The SD values of the two-split quantizers and the three-split 1-DP-SVQs (1-DP-3-SVQs) are presented in Figure 4.15 for the validation database and in Figure 4.16 for the training database. As in the case of the three-split quantizers, again IS-SVQs achieve the best results. The 22-bit 2-IS-SVQ is relatively close to the limits of transparent quality. The performance of the 22-bit 2-IS-SVQs and $DP_{IS\text{-}641}$-SVQ are essentially the same. The effectiveness of the two-split quantizers compared with the three-split quantizers can clearly be observed, since the two-split 1-DP-SVQs have noticeably lower SD values than its three-split counterpart. However the performance of the two-split quantizers drops considerably in the validation data compared to the training data. With the memoryless two-split quantizers (2-SVQ) the SD_{ave} values increase about 0.2 dB in the validation database compared to the training database. With 2-IS-SVQs the corresponding number is 0.1 dB.

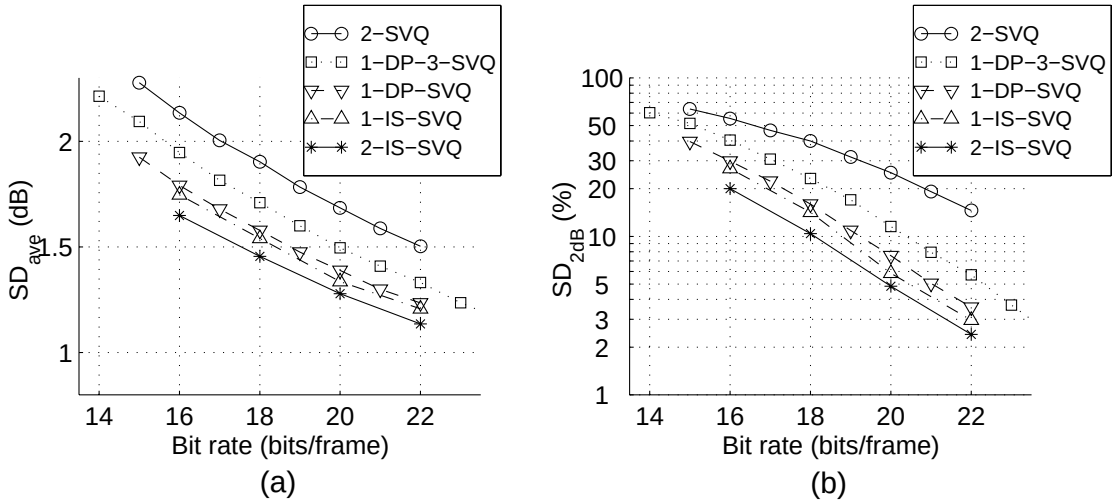

Figure 4.15. Comparison of SD values of the two-split quantizers for the validation database. The three-split 1-DP-SVQ is denoted with 1-DP-3-SVQ.

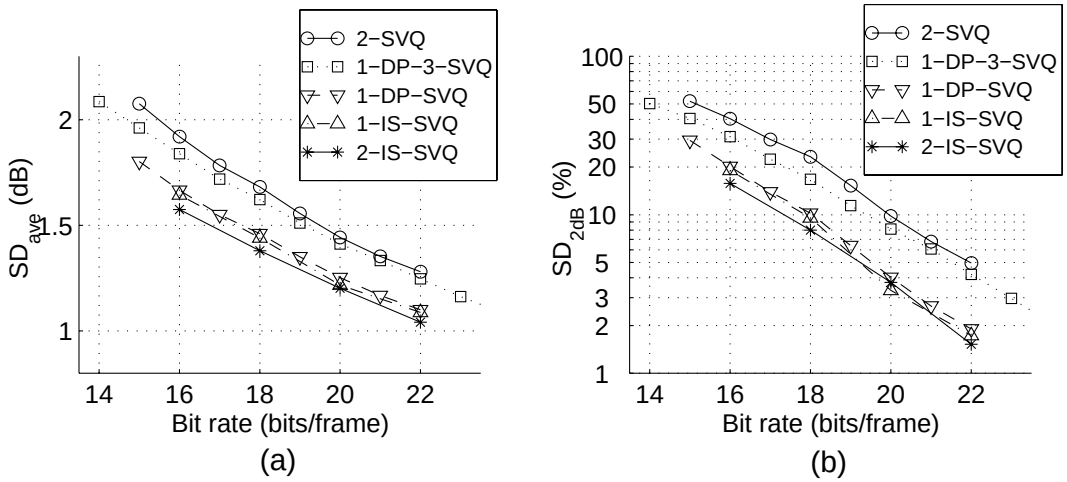

Figure 4.16. Comparison of SD values of the two-split quantizers for the training database. The three-split 1-DP-SVQ is denoted with 1-DP-3-SVQ.

Table 4.6 presents the relative bit rate estimates of the quantizers compared to 1-DP-SVQs. Similar conclusions can be made than in the three-split quantizer case. Nevertheless IS-SVQs are not as efficient compared to 1-DP-SVQs in thet two-split case than in the three-split case. Here, the advantage of using 2-IS-SVQ instead of 1-DP-SVQ is about one bit. In addition, IS-SVQs outperform DP-SVQs and FP-SVQs having the same prediction order, on average by 0.2–0.4 bits. As in the three-split case, the relative bit rate estimates of DP-SVQs and FP-SVQ are nearly equivalent. Finally, the table shows that a bit save of about 2.5 bits can be achieved using 2-IS-SVQs instead of 1-DP-3-SVQs.

Table 4.6. The relative bit rate estimates of 15–22-bit two-split quantizers compared to 1-DP-SVQs. The estimates are calculated from the validation and the training databases.

Quantizer	*Rel. bit rate (valid.)*	*Rel. bit rate (train.)*
2-SVQ	-3.27	-2.22
1-DP-3-SVQ	-1.42	-1.78
1-DP-SVQ	0.00	0.00
1-FP-SVQ	0.02	0.03
1-IS-SVQ	0.45	0.25
2-DP-SVQ	0.86	0.56
2-FP-SVQ	0.93	0.63
2-IS-SVQ	1.31	0.75
3-DP-SVQ	1.15	0.77
3-FP-SVQ	1.27	0.87

4.5 Conclusions

4.5.1 Summary

In previous sections the objective results of quantizers were shown. The two-split and the three-split quantizers using either diagonal matrix, full matrix or inter-split predictors were evaluated. It was noticed that a substantial gain can be achieved using advanced predictive structures, like inter-split predictor, in quantizers. In the three-split case, IS-SVQs outperform DP-SVQs and FP-SVQs with 0.7–0.8 bits. However in the two-split case, the corresponding gain is only 0.2–0.4 bits. The difference between the performance of DP-SVQs and FP-SVQs is practically nonexistent.

The two-split quantizers gain 0.4–2 bits compared to the three-split quantizers. The gain is most evident in memoryless quantizers and in the training database. However the advantage diminishes in the validation database and in sophisticated quantizers. For example the two-split 2-IS-SVQs outperform the three-split 2-IS-SVQs only with 0.4 bits in the validation database.

The order of predictor has distinct influence to the efficiency of the quantizer. The memoryless quantizer requires more than 2 bits in the training database and 3 bits in the validation database to achieve similar performance than simplest predictive quantizer 1-DP-SVQ. Furthermore the gain of using second order predictors instead first order is about 0.6–0.8 bits. Thus predictive quantizers are inevitable for low bit rate coders. Although the error robustness declines because of a predictor, the error propagation can be limited to the number of frames defined by the order of the predictor. For this reason lower order predictors are recommendable than higher order predictors. Since the gain achieved by third order predictors compared to the second order predictors is small a reasonable upper limit to the order of predictor is two.

Some observations can be made from optimal bit allocation of sub-quantizers of splits. In the three-split quantizers the sub-quantizer of first split gets by at least one bit lower than the sub-quantizers of other splits. In the two-split quantizers the first split gets by even more than one bit lower than second split. Nevertheless the optimal allocation must be found experimentally.

The limits of transparent quality by Paliwal and Atal can be achieved with several the three-split quantizers. A quantizer of the lowest bit rate that fulfills the limits both in the training and validation database is 23-bit 2-IS-3-SVQ. There are several 24-bit quantizers which achieve the limits, for example 1-IS-3-SVQ, 2-DP-3-SVQ and 2-FP-3-SVQ. The best two-split quantizer, 22-bit 2-IS-SVQ, remains slightly from the limits, since SD_{ave} equals to 1.14 dB and SD_{2dB} equals to 2.7 % in the validation database.

The performance of $DP_{IS\text{-}641}$-SVQ depends significantly on testing data properties. For example $DP_{IS\text{-}641}$-SVQ achieves poor SD values in the unfiltered part of the training database. Nevertheless when the same data is filtered with modified IRS-filter, then $DP_{IS\text{-}641}$-SVQ gains markedly better. Anyway the performance of $DP_{IS\text{-}641}$-SVQ can be reached with any of the 25-bit predictive quantizers presented in previous sections. In addition several quantizers overtake the performance of $DP_{IS\text{-}641}$-SVQ with even lower bit rates, for example the 23-bit 2-IS-3-SVQ and the 24-bit 1-IS-3-SVQ. Even the 22-bit 2-IS-2-SVQ obtains the same performance.

As it has already been mentioned the validation database has been differently prefiltered than the training database. It can be said from the results that sophisticated predictive quantizers tolerate better unsuitable data than simple predictive or memoryless quantizers. In addition the three-split quantizers get over the validation database better than the two-split quantizers.

4.5.2 Predictor and prediction coefficients

The predictive quantizers described in this thesis have better quality than $DP_{IS\text{-}641}$-SVQ. This is caused by better predictors. The prediction coefficients of $\mathbf{B}_1$ of $DP_{IS\text{-}641}$-SVQ have values below 0.5, exactly the diagonal coefficients of $\mathbf{B}_1$ are (0.29, 0.33, 0.38, 0.41, 0.44, 0.36, 0.32, 0.30, 0.26, 0.20). These values are much lower than typical diagonal values of $\mathbf{B}_1$ in this thesis. For example, the 25-bit 1-DP-3-SVQ that has similar predictor than $DP_{IS\text{-}641}$-SVQ, has corresponding coefficients (0.53, 0.43, 0.50, 0.51, 0.55, 0.56, 0.60, 0.53, 0.53, 0.53). Since the prediction parameters are higher the predictor works effectively making the quantization process accurately. The inter-split predictor is even more powerful than diagonal matrix predictor. As an example prediction coefficient matrices, $\mathbf{B}_0$ and $\mathbf{B}_1$ of the 24-bit 1-IS_{321}-SVQ are presented:

$$\mathbf{B}_0 = \begin{bmatrix} 1 & 0 & 0 & 0.09 & 0.02 & -0.03 & 0.05 & -0.04 & 0.04 & -0.17 \\ 0 & 1 & 0 & 0.10 & 0.07 & -0.06 & 0.07 & -0.03 & 0.06 & -0.19 \\ 0 & 0 & 1 & 0.44 & -0.04 & 0.09 & 0.13 & 0.08 & 0.08 & -0.16 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0.26 & -0.05 & 0.18 & -0.23 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0.34 & 0.08 & 0.06 & -0.19 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.55 & 0.01 & 0.16 & -0.20 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad (4.1)$$

$$\mathbf{B}_1 = \begin{bmatrix} 0.56 & 0.00 & 0.05 & 0.07 & 0.01 & -0.01 & 0.04 & -0.01 & 0.03 & -0.12 \\ 0.24 & 0.30 & 0.07 & 0.08 & 0.04 & -0.03 & 0.04 & -0.01 & 0.04 & -0.11 \\ 0.17 & 0.02 & 0.46 & 0.27 & -0.05 & 0.10 & 0.07 & 0.08 & 0.04 & -0.03 \\ 0.08 & 0.01 & 0.10 & 0.61 & 0.08 & 0.07 & 0.17 & -0.03 & 0.12 & -0.08 \\ 0.00 & 0.06 & -0.04 & 0.04 & 0.65 & 0.13 & 0.25 & 0.05 & 0.04 & -0.11 \\ 0.02 & -0.01 & 0.01 & -0.03 & 0.08 & 0.65 & 0.43 & 0.01 & 0.14 & -0.12 \\ 0.06 & -0.01 & -0.03 & 0.01 & 0.00 & 0.08 & 0.68 & 0.06 & 0.07 & 0.02 \\ -0.01 & 0.01 & 0.01 & -0.02 & 0.03 & -0.01 & 0.08 & 0.58 & 0.10 & 0.01 \\ -0.01 & -0.01 & 0.00 & 0.02 & -0.03 & 0.03 & 0.02 & 0.02 & 0.57 & 0.11 \\ -0.08 & 0.02 & 0.01 & 0.00 & -0.02 & -0.01 & 0.01 & 0.01 & 0.04 & 0.59 \end{bmatrix} \quad (4.2)$$

As it can be seen the elements near the diagonal of the matrices are relatively high. There are nine elements having value above 0.5 and 20 elements having absolute value between 0.1 and 0.5.

4.5.3 Alternative LSF quantizers for IS-641

Finally three consistent quantizers for $DP_{IS\text{-}641}$-SVQ are presented. All these three have similar performance than $DP_{IS\text{-}641}$-SVQ. These quantizers are 22-bit 2-IS_{21}-SVQ, 23-bit 2-IS_{123}-SVQ and 24-bit 1-IS_{321}-SVQ. Their performance is presented in Table 4.7 for the validation database and in Table 4.8 for the training database.

The tables show that up to four bits can saved using 2-IS-2-SVQ instead of $DP_{IS\text{-}641}$-SVQ. The 2-IS-3-SVQ and 2-IS-3-SVQ do not only outperform $DP_{IS\text{-}641}$-SVQ but also fulfill the requirements of transparent quality.

However the bit rate and the spectral distortion are not the only measure to evaluate LP quantizer. The other important measures are complexity of computation and memory requirements. Table 4.9 presents worst case numbers of additions and multiplications per frame (20 ms) and memory requirements of the quantizers. Depending on implementation the calculations can be either integer or floating point operations. Here the weighting matrix from Equation (2.12) is assumed to be used, thus nearly doubling the need of operations. The table shows that 2-IS-2-SVQ requires nearly 58000 operations per frame, that is 2.9 million operations per second. This is quite demanding compared to the other quantizers. On the other hand, the 23-bit 2-IS-3-SVQ need more than ten times fewer operations and memory than 2-IS-2-SVQ and nearly half of the demand of $DP_{IS\text{-}641}$-SVQ. However the number of operations can be reduced considerably using sophisticated searching algorithms.

Table 4.7. Objective measures of the quantizers having similar performance than IS-641 in the validation database. The bit allocation of the quantizers are shown in Alloc.-column. The number of unstable filters is denoted by Unst.

Quantizer	Alloc.	SD_{ave} (dB)	SD_{2dB} (%)	SD_{4dB} (%)	Unst.	d_{AQ} (Hz)	d_{AQ1} (Hz)	d_{AQ2} (Hz)	d_{AQ3} (Hz)
2-IS-2-SVQ	10,12	1.14	2.4	0.0	13	26.3	20.5	28.7	-
2-IS-3-SVQ	7,8,8	1.11	2.0	0.0	7	25.3	21.9	22.4	26.7
1-IS-3-SVQ	7,8,9	1.09	1.6	0.0	4	24.9	20.9	24.1	25.4
$DP_{IS\text{-}641}$-SVQ	8,9,9	1.10	1.8	0.0	5	25.4	20.3	22.5	28.0

Table 4.8. Objective measures of the quantizers having similar performance than IS-641 in the training database. In last rows the results of $DP_{IS\text{-}641}$-SVQ are presented for modified IRS (modIRS) and flat (flat) filtered parts of the database. The number of unstable filters is denoted by Unst.

Quantizer	Alloc.	SD_{ave} (%)	SD_{2dB} (%)	SD_{4dB} (%)	Unst.	d_{AQ} (Hz)	d_{AQ1} (Hz)	d_{AQ2} (Hz)	d_{AQ3} (Hz)
2-IS-2-SVQ	10,12	1.04	1.5	0.0	49	25.0	19.4	27.2	-
2-IS-3-SVQ	7,8,8	1.06	2.1	0.0	59	25.5	21.1	23.7	26.5
1-IS-3-SVQ	7,8,9	1.04	1.5	0.0	33	24.8	20.0	25.4	24.6
$DP_{IS\text{-}641}$-SVQ	8,9,9	1.15	3.7	0.0	36	27.4	23.0	24.8	28.5
$DP_{IS\text{-}641}$-SVQ (modIRS)	8,9,9	1.05	1.3	0.0	17	24.8	19.8	23.2	26.4
$DP_{IS\text{-}641}$-SVQ (flat)	8,9,9	1.24	6.1	0.0	19	30.0	26.2	26.4	30.5

Table 4.9. Computation and memory requirements of quantizers having similar performance than IS-641. The third column shows the total number of additions and multiplications per frame in the worst case. The fourth column shows the number of the codebook and the prediction parameters needed to store.

Quantizer	*Bits*	*Operations*	*Memory*
2-IS-2-SVQ	22	57568	28896
2-IS-3-SVQ	23	4585	2409
1-IS-3-SVQ	24	6533	3333
$DP_{IS\text{-}641}$-SVQ	26	8714	4362

5 Subjective Tests

To evaluate the perceptual quality of the LSF quantizers, two subjective *Degradation Category Rating* (DCR) [49] listening tests were carried out. The purposes of the first test are, firstly, to discover the relation of a subjective speech quality compared to a bit rate used for the quantization and, secondly, to compare the perceptual quality and objective quality measures of quantization. In the second test, speech frames are classified either to voiced or unvoiced sounds and the perceptual qualities of LSF quantization of these classes are under the study. In addition the test examines the effect of reduced excitation quality of a residy signal to the required LSF quantization quality.

It has been noticed that the DCR procedure affords high sensitivity in distinguishing among good quality test samples. Thus the procedure is feasible and suitable to separate relatively tiny differences of LSF quantizers. In the DCR test, pairs of sentences are listened, where the first sentence is a reference and the second one is the same sample processed by the system under evaluation. The purpose of the reference sample is to anchor each judgment of the listeners. The listeners judge the degradation with a five point scale, presented in Table 5.1. Numbers of sentences and speakers are used to avoid the effect of the error, however so that the set of the test samples is same for every candidate. It is recommended that at least 12 listeners should participate in the test to achieve statistical reliability of results. The result of the DCR test is called Degradation Mean Opinion Score (DMOS). Beside the processed samples, Modulate Noise Reference Unit (MNRU) samples can be included to the test to achieve comparability to other listening tests. The DCR procedure is described in detail in ITU-T Recommendation P.80 [49].

Table 5.1. Degradation category scale for the DCR listening test.

Score	Subjective judgment
5	Degradation is inaudible
4	Degradation is audible but not annoying
3	Degradation is slightly annoying
2	Degradation is annoying
1	Degradation is very annoying

In the tests, the input speech material consisted of six British-English speech samples (three males and three females) from NTT's *Multi-Lingual Speech Database for Telephonemetry 1994* speech database [21]. Each sample contains one sentence. The lengths of the samples range from 2.5 s to 3.5 s producing 18 s in total, that is, 900 frames. The samples contain 600 frames active speech, corresponding to 67 % of the material. Following to ITU-T recommendation, speech samples are downsampled to 8 kHz, modified IRS filtered, processed and upsampled back to 16 kHz. The tests are performed by computer program and samples are listened with high-quality headphones. The bandwidth of the signal is limited to be between 100 Hz and 3600 Hz by a high quality filter. A special attention has been devoted to the listening environment, to minimize the effect of noise and other disturbing elements.

In both tests the quantizers of bit rates 14, 17, 20, 23 and 26 bits per frame are used. The bit rates are chosen to contain both good and poor quality quantizers. The quantizers are implemented to the IS-641 codec. Used quantizers have diagonal matrix predictors of order one, like IS-641. The details of the 1-DP-SVQs can be found in Section 4.3.

To compare DMOS values to objective quality measures of the quantization, spectral distortion values (SD_{ave} and SD_{2dB}) and segmental signal-to-noise ratio (segSNR) are calculated for each sample (see Section 2.5). Because of the large amount of data, only the most important results are presented in this chapter and detailed results of the tests can be found in Appendices I and II.

5.1 Degradation Category Rating of LSF Quantizers

In this section the basic relations between subjective speech quality and bit rate used for the LSF quantization are examined. Along with the 14–26-bit quantizers, also the original IS-641 codec with (IS) and without (NQ) LSF quantization are included in the test. Beside the samples processed with these codecs, 5, 15 and 30 dB MNRU reference samples and original clean source are included in the test.

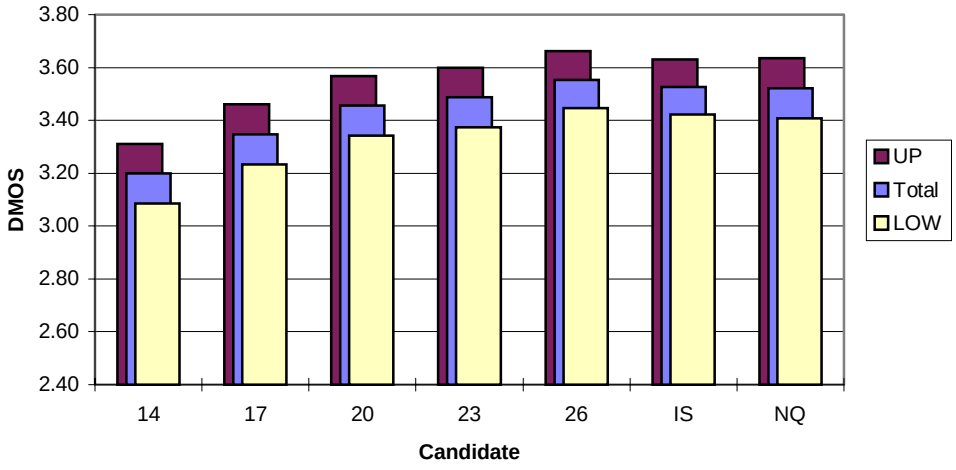

Figure 5.1. DCR listening test results of the LSF quantization. Total is the average value of the results, UP is the upper 95% confidence limit and LOW is the lower 95% confidence limit.

The listening test was carried out in the Speech and Audio Systems Laboratory at the Nokia Research Center and in the Laboratory of Acoustics and Audio Signal Processing in the Helsinki University of Technology. A test environment has been built as similar as possible in both places and dependencies of place or environment can not be noticed from the listening test results. The results of 37 expert listeners were evaluated. The results of the test are presented in Figure 5.1 and Table 5.2.

The quantizer of 14 bits per frame obtained the DMOS value 3.20 compared to the original unprocessed speech, and the score tends to 3.5 as the bit rate increases. However, the quantizers having a bit rate 20 bits per frame or higher do not have statistically significant differences in their DMOS values, because the 95% confidence limits differ ±0.1 DMOS from the mean value. Consequently, the perceptual quality of the LSF quantization seems to saturate for a bit rates over 20 bits in the IS-641 speech codec. According to this test, the bit rate of LSF quantization could be decreased from original 26 bits to 20 bits without any affect on perceptual quality of synthesized speech. Although the number of listeners is sufficient, the number and types of test samples are limited. Therefore the result applies only to the same type of speech material used in the test. Thorough validation tests with different background noise conditions, speech samples, input levels and tandem conditions should be done, before decreasing the actual bit rate. Another reason for such result can be the IS-641 coder itself. The perceptual quality of coder depends on a number of parameters of the coder. Thus in certain situations a “bottleneck” parameter can settle the upper limit to the perceptual quality, since a slightly decreased quality or bit rate of other components does not affect the overall quality.

Table 5.2. Listening test results of different bit rates in LSF quantization. In columns 2 to 7 the DMOS values of different samples are presented; Total is an average value of the results, STD is a standard deviation, UP is an upper 95% confidence limit and LOW is a lower 95% confidence limit.

	F1	F2	F3	M1	M2	M3	**Total**	STD	UP	LOW
14	3.08	2.89	2.92	3.65	2.78	3.86	**3.20**	0.057	3.31	3.09
17	2.97	2.89	3.14	3.86	3.08	4.14	**3.35**	0.059	3.46	3.23
20	3.08	3.11	3.22	3.84	3.27	4.22	**3.45**	0.058	3.55	3.34
23	3.22	3.05	3.30	3.89	3.08	4.38	**3.49**	0.057	3.60	3.37
26	3.43	3.30	3.24	3.89	3.27	4.19	**3.55**	0.055	3.66	3.45
IS	3.32	3.08	3.41	3.97	3.32	4.05	**3.53**	0.053	3.63	3.42
NQ	3.03	3.27	3.30	4.08	3.24	4.22	**3.52**	0.058	3.64	3.41
MNRU 5	1.24	1.14	1.05	1.30	1.08	1.30	**1.18**	0.027	1.24	1.13
MNRU 15	2.14	2.00	1.76	2.95	1.95	2.19	**2.16**	0.052	2.26	2.06
MNRU 30	4.51	4.24	4.35	4.65	4.51	4.59	**4.48**	0.042	4.56	4.39
Direct	4.86	4.78	4.81	4.76	4.92	5.00	**4.86**	0.030	4.91	4.80

Figures 5.2 and 5.3 visualize the DMOS values of the individual samples. The figures and Table 5.2 show that the scores of the quantizers are very case dependent. Although the quantizers seem to get illogical scores, when separate samples are compared, the total results are rather consistent. A decrease of DMOS value when the bit rate increases can be explained by statistical inaccuracy and variations in the operation of the quantizers. The 5 dB MNRU value sets the lowest limit to the scores and higher MNRU values are ranked above that limit. It is likely that inconsistencies of average DMOS values diminish in a more extensive test.

A noteworthy detail is that male voices M1 and M3 obtain on average 0.5 better results than the rest of the speakers. There can be several technical and psychological reasons for this. The first reason is a low fundamental frequency of male voice, that is, a long pitch lag. It effects to the residual signal so that no more than one pitch period is in a subframe. Therefore the excitation signal is easy to code with few pulses in IS-641 codec. This is a consequence of a general fact that the CELP codecs, which are working in a time scale, work effectively with low male voices. Still the MNRU reference samples indicate that the type of codec is not the only reason. Another reason could be the psychological effect of reliability and security of low male voice that might reflect to the pleasantness of voice. Unfortunately within this thesis there is no opportunity to examine these interesting assumptions more accurately.

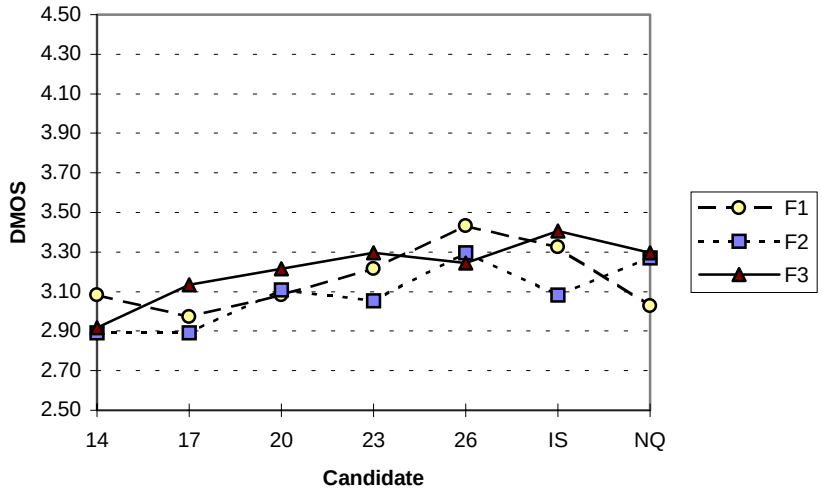

Figure 5.2. Test results of female voices in LSF quantization.

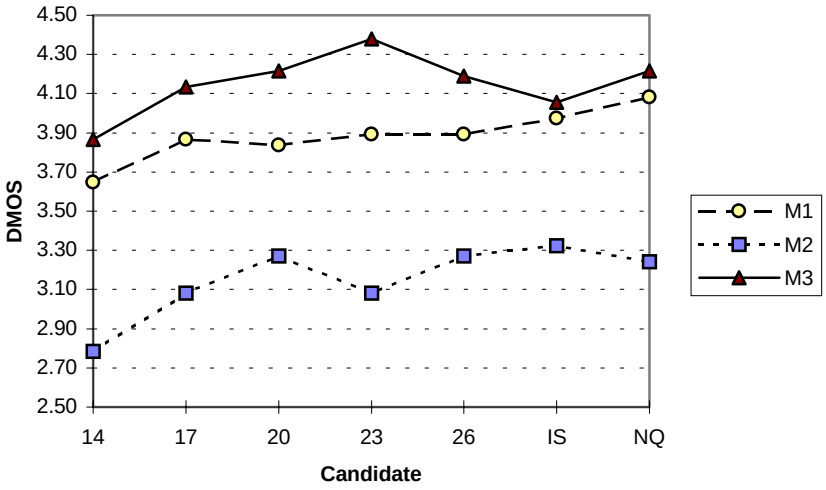

Figure 5.3. Test results of male voices in LSF quantization.

The SD and the corresponding DMOS values of speech samples are presented in Figures 5.4(a) and 5.4(b). Naturally the unquantized codec (NQ) yields zero spectral distortion. The maximum SD_{ave} value is 2.2 dB and the maximum SD_{2dB} value is 57%, both in the 14-bit codec. The quantizer of the IS-641 codec, with SD_{ave} value equal to 1.1 dB and SD_{2dB} value equal to 2.3 %, stays slightly behind from the 26-bit quantizer. Figures 5.4(a) and 5.4(b) show that the SD_{ave} and logarithm of the SD_{2dB} values are corresponding to the DMOS values. From Figure 5.4(d) it can be seen that the objective SD_{ave} and segSNR values correspond. Within these results, accurate objective limits of perceptually sufficient quantization level of LSF parameters are infeasible to define.

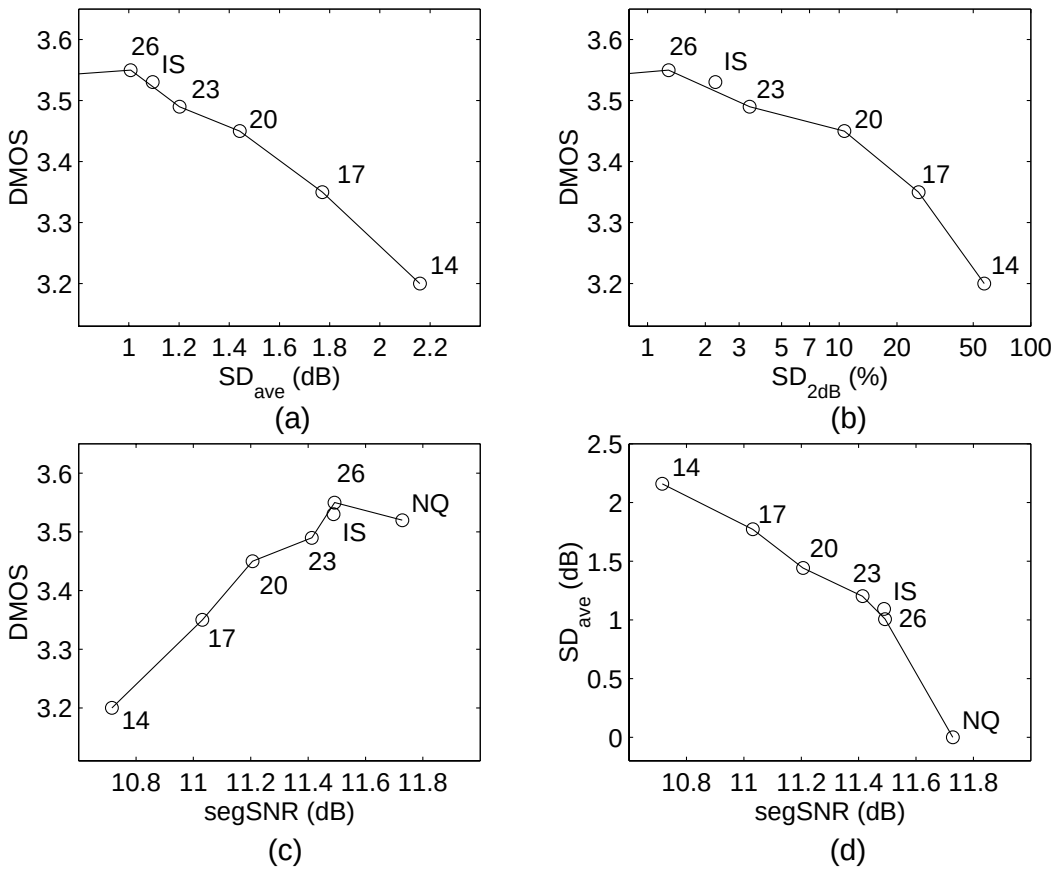

Figure 5.4. Comparison of different quality meters of LSF quantization. Logarithmical x-scale is used in (b). In (a) and (b), the candidate NQ does not lie on the visible graph area.

Paliwal and Atal have proposed that to achieve transparent quantization of LPC parameters, the SD_{ave} should be about 1 dB and SD_{2dB} should be less than 2% [9]. Only the 26-bit quantizer, with SD_{ave} value equal to 1.0 dB and SD_{2dB} value equal to 1.3%, fulfills the requirement. However, the DMOS values prove that there are no significant statistical differences between codecs from unquantized to a 20-bit codec. The fact that the 20-bit quantizer obtains SD_{ave} value 1.4 dB and SD_{2dB} value 11 % indicates that the SD limits can be loosened in certain situations. In addition Paliwal and Atal note that transparent quantization of LPC parameters may be possible with higher percentage of outlier frames, but they have not investigated it. In fact, they do not present a validation subjective test of the transparent quality limit of SD_{ave} either. Altogether, it seems that the subjective and the objective measures correspond in some sense, but there is needless margin between current and perceptually sufficient limits.

5.2 Relation of voiced and unvoiced frames LSF quantization and effect of excitation signal quality

This test is divided into two parts. In the first part, the purpose is to compare different excitation and LSF quantization qualities in the speech codec and to discover their relation to the overall quality. In the second part the speech frames are first labeled as voiced or unvoiced sounds. Then a perceptual quality of LSF quantization according to these classes is to be examined. However the speech samples of both parts were listened and evaluated at the same time. Beside the samples processed by codecs with 14–26-bit 1-DP-SVQs, test material includes 10, 20 and 30 dB MNRU reference samples and original clean sources. The listening test was carried out in the Speech and Audio Systems Laboratory at the Nokia Research Center. The results of 24 expert listeners were evaluated.

5.2.1 Effect of excitation signal quality

In the first part of the test, the subjective quality of codecs with different excitation qualities is to be examined. Consequently the residual signal is coded either with four excitation pulses in subframe, as in IS-641, or with reduced three pulses. The motivation to do this is that the bit rate needed to code excitation pulses can be reduced considerably removing by one pulse. The original IS-641 codec uses 17 bits per subframe, that is 3400 bits per second, to code excitation positions and signs of four excitation pulses, while the three- pulse version uses only 12 bits per subframe. Therefore the total bit saving attained by the three-pulse version is 20 bits per frame or 1000 bits per second. From here on the codecs are noted in a format (number of pulses, bit rate of LSF quantizers). For example, three- pulse and 20-bit quantizer codec would be noted (3, 20). In following two paragraphs the bit allocations of the three- and four-excitation-pulse version are presented.

The bits for the four excitation pulses are divided as follows. All pulses can have the amplitude +1 or –1. The positions of pulses in a subframe are divided into four tracks, where each track contains one pulse, as shown in Table 5.3. Therefore the bits are shared between signs; four bits, and positions; 3 + 3 + 3 + 4 bits, thus making total of 17 bits per subframe.

The three excitation pulses can also have the amplitude +1 or –1. The positions of the pulses are shown in Table 5.4. Besides this, all the pulses can simultaneously be shifted by one position (+1) to occupy the odd pulse positions. The total of 12 bits per subframe contains three bits for signs and 2 + 3 + 3 bits for the positions and one bit for shifting. In both cases closed-loop search algorithm is used to find the best possible places and signs for pulses.

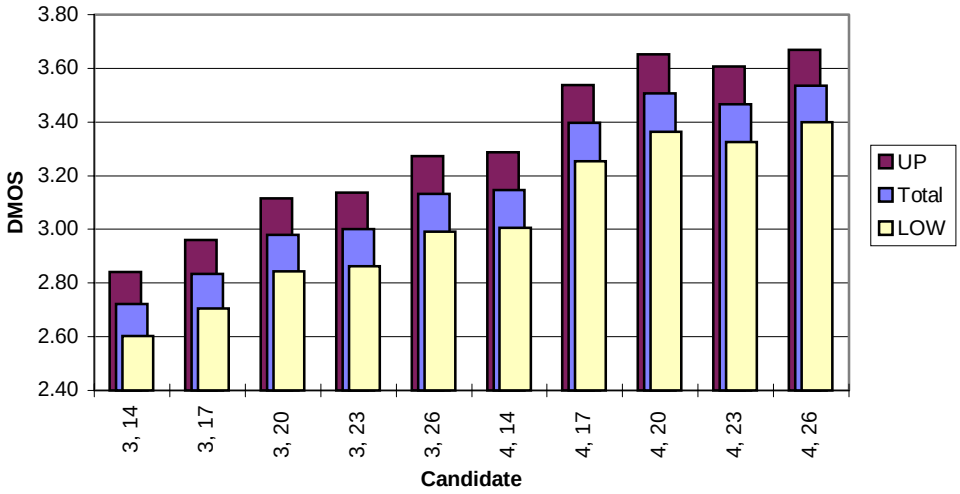

Figure 5.5. Listening test results of three- and four-pulse codecs.

Table 5.3. Potential positions of four excitation pulses.

Pulse	Positions
i_0	0, 5, 10, ..., 35
i_1	1, 6, 11, ..., 36
i_2	2, 7, 12, ..., 37
i_3	3, 8, 13, ..., 38, 4, 9, 14, ..., 39

Table 5.4. Potential positions of three excitation pulses.

Pulse	Positions
i_0	0, 10, 20, 35
i_1	2, 12, 22, 32, 4, 14, 24, 34
i_2	6, 16, 26, 36, 8, 18, 28, 38

The listening test results of the three- and four-pulse versions of the codecs are shown in Figure 5.5. The quality of the codec declines considerably along the reduction of the excitation quality. The (4, 14) codec obtains about the same DMOS value 3.1 as the (3, 26) codec. The quality of the three-pulse version decreases from DMOS value 3.13 of 26-bit quantizer to the DMOS value 2.72 of the 14-bit quantizer, making a total of 0.41 DMOS degradation while the four-pulse version makes 0.38 DMOS degradation. Though the quality decreasing of the three-pulse version is almost linear, the corresponding saddle point of (4, 20) codec exists. The DMOS values of the four-pulse version codec follow closely the results of the first listening test, and all 20–26-bit quantizers get a score around 3.5. The results of four-pulse version are consistent with the first listening test results.

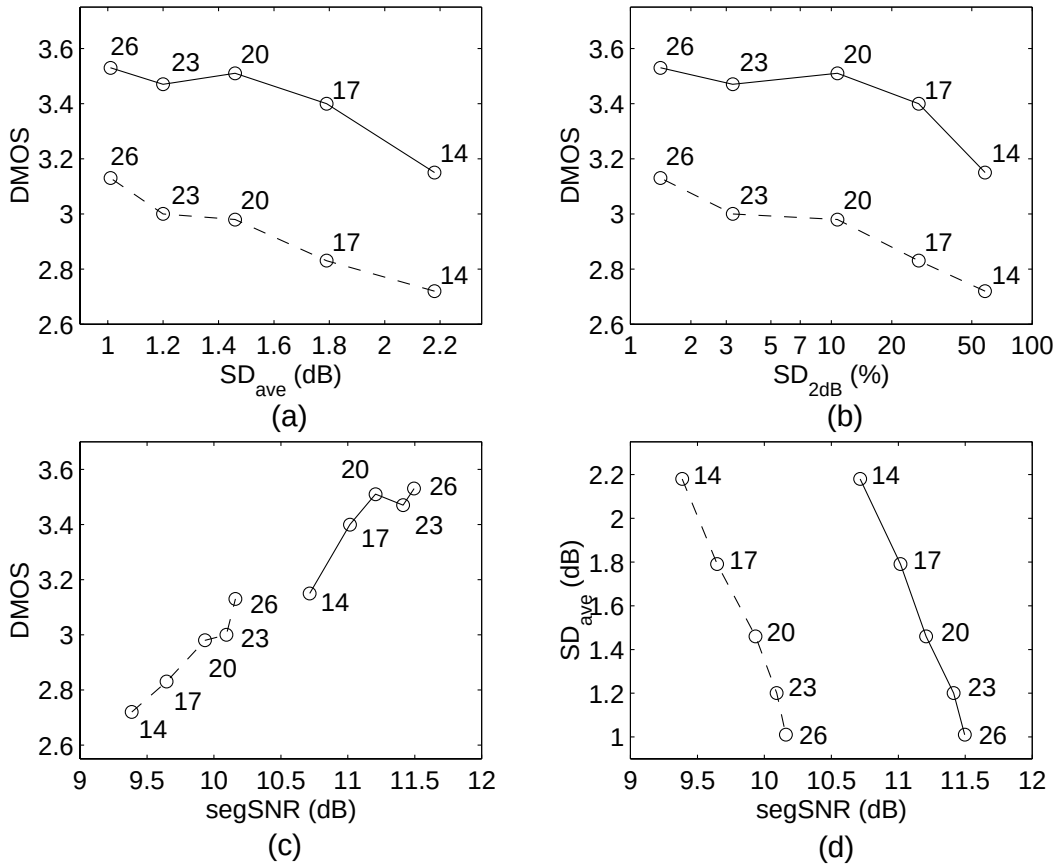

Figure 5.6. Comparison of different quality meters of LSF quantization. Dashed line: three-excitation- pulse codec, and solid line: four-excitation-pulse codec. Logarithmic x-scale is used in (b).

Next the DMOS values are compared with the numerical measures of LSF quantization. The SD and the segSNR values are visualized in Figure 5.6. In Figures 5.6(a) and 5.6(b) it can be seen that the saddle point of the line of the four-pulse codec has almost disappeared from the three-pulse version. It seems that the poor quality of excitation favors an improvement in the quantization of LPC parameters to come audible while contrary an enhanced excitation quality makes the quantization quality improvement inaudible after a certain saddle point. Thus the quality of LPC parameters and excitation pulses have a dependence: if one performs weakly, the other one can compensate the overall result. Nevertheless the overall quality appears to saturate with good-quality excitation pulses and LSF quantizers.

The most interesting result is presented in Figure 5.6(c). The three- and four-pulse versions of codec have 0.55 dB difference in segSNR values but the same DMOS. It appears that the degradation in segSNR scale is not comparable to the DMOS scale. This indicates that segSNR is an applicable measure only to compare same types of changes in the codec. However, as an LSF quantization measure in CELP style of codec, the segSNR values have similar performance as SD values.

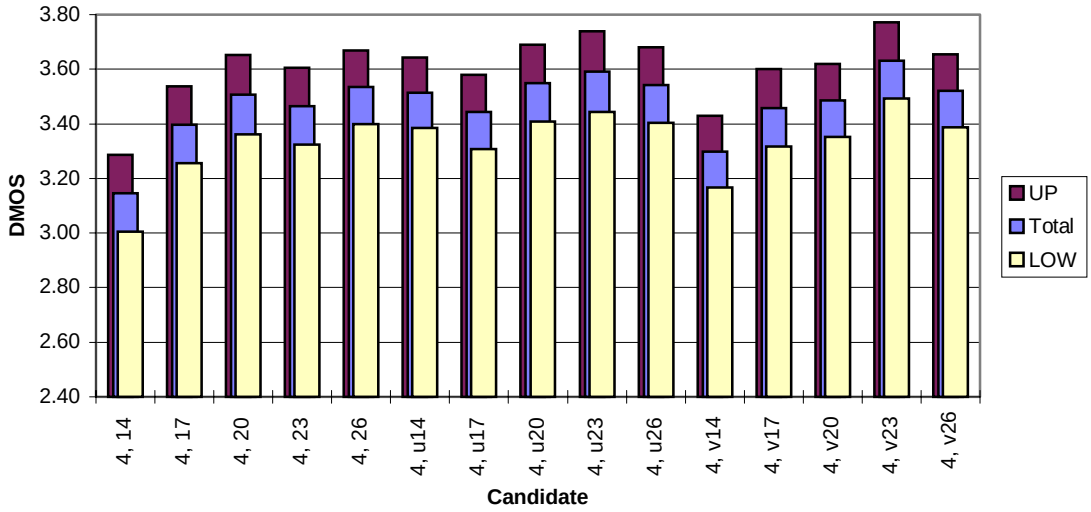

Figure 5.7. The listening test results of LSF quantization in the unvoiced (u) and the voiced (v) frames.

5.2.2 Relation of voiced and unvoiced frames LSF quantization

The results of the LSF quantization of voiced and unvoiced frames are presented in Figure 5.7. All these codecs utilize four excitation pulses as IS-641. In the voiced quantization case (v) the LSF parameters of voiced frames are quantized, while unvoiced remain unquantized, and on the contrary in the unvoiced case (u). The classification to voiced and unvoiced frames was made firstly by the decision based on pitch correlation energy. Later the material was examined and corrected manually. The active speech, used in this DCR test, contains 73 % voiced and 27 % unvoiced frames.

Compared to the codecs that quantize LSF parameters in every frame, the unvoiced- and voiced-frame LSF quantization codecs have higher average DMOS values, respectively 0.11 and 0.07 DMOS. This can be expected, because there are unquantized frames in both later cases. However, the trend of DMOS values is exciting. The DMOS values of voiced LSF quantization codecs decrease with bit reduction, while the DMOS values of unvoiced LSF quantizer codecs remain around 3.5 regardless of bit reduction. This perceptual effect on the accuracy of hearing error in spectra of voiced sound and on the contrary, tolerance of error in spectra of unvoiced sound has been presented in literature earlier [50, 51, 52]. Hagen *et al.* [53] have proposed that different SD criteria should be used for unvoiced and voiced frames. They based the study on a pair comparison listening test. They proposed the “transparency” rule for the quantization of unvoiced spectra in CELP codec to be; the average SD must be at most 2.0–2.1 dB, and the percentage of frames having SD above 4 dB must be less than 1%. Their rule of transparency for quantization of voiced spectra

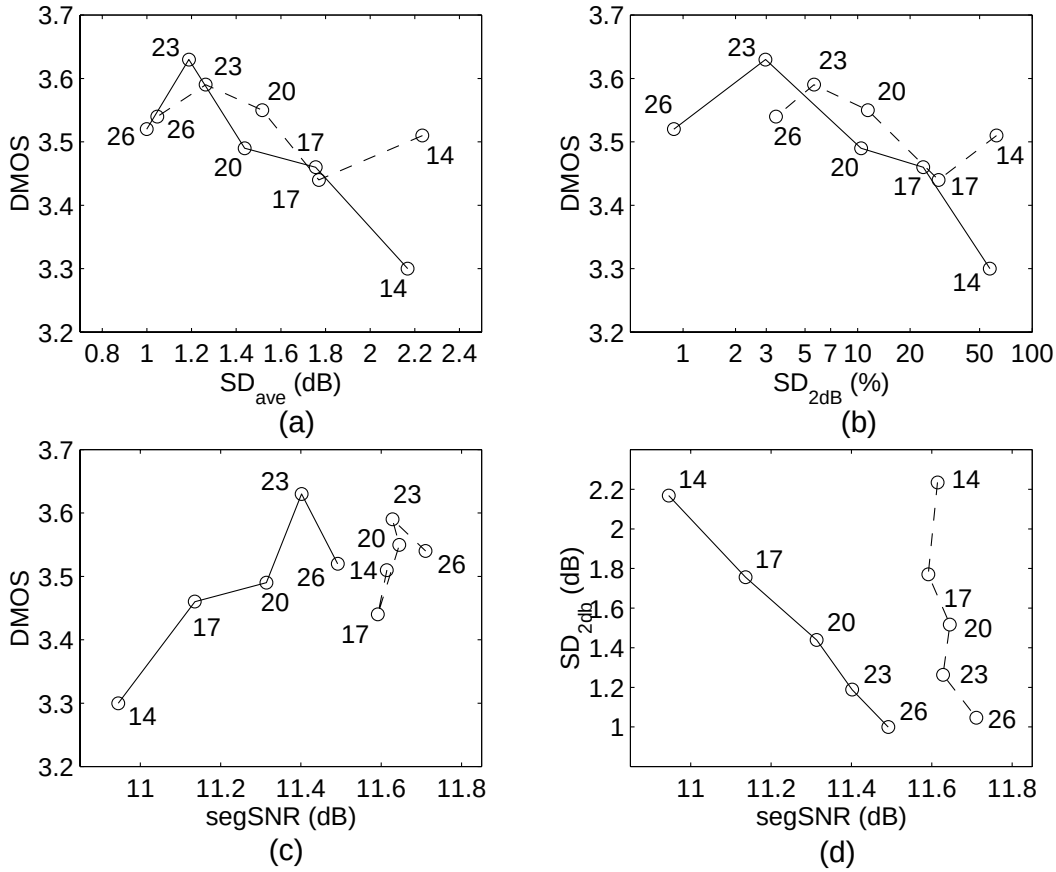

Figure 5.8. Quantization of LSF parameters in voiced (solid line) and in unvoiced frames (dashed line). Logarithmical x-scale is used in (b).

followed the criteria of Paliwal and Atal [9] presented earlier, with exception of tightened 1% limit to outlier frames having SD over 2 dB.

Following the rules proposed by Hagen *et al.*, the SD values for unvoiced frames were calculated. The SD and corresponding DMOS values are shown in Figure 5.8 with dashed line. Up to 17-bit quantizer with SD_{ave} value equal to 1.5 dB and SD_{4dB} value equal to 0%, their rule for transparent quantization of spectra is fulfilled. The 14-bit quantizer has the SD_{ave} value 2.2, which is slightly over limits proposed by Hagen *et al.* Since the DMOS values of the quantizers are relatively similar, the result of the test support relieved criteria of the SD values for the unvoiced spectra by Hagen *et al*. The segSNR values of unvoiced frames have very small correlation with DMOS values in Figures 5.8(c) and 5.8(d). It seems that neither segSNR nor SD values are accurate for discriminating the perceptual qualities of unvoiced sounds.

Solid lines in Figure 5.8(a) and 5.8(b) present the SD and the DMOS values of the voiced frames. Only the 26-bit quantizers, with SD_{ave} value 1.00 dB and SD_{2dB} value 0.89 %, fulfill the requirements of Paliwal and Atal, and Hagen *et al*. The 23-bit quantizer lacks slightly

the requirement of Paliwal and Atal. As earlier assumed, there appears to be purposeless margin between recommended and perceptually satisfactory limits.

5.3 Conclusions

The perceptual qualities of the basic moving average LSF quantizers were evaluated. The results point out that the quantizers designed by the method presented in this thesis have a slightly better or similar subjective performance compared to the quantizer of the IS-641 codec. The perceptual quality of 26-bit quantizer corresponds to original quantizer of IS-641. Within these tests the qualities of 20- and 23-bit quantizers were also sufficient compared to IS-641.

The results indicate that the DMOS values and the numerical measures, SD and segSNR, correspond to each other, but accurate objective limits of the perceptually sufficient quantization level of LSF parameters are difficult to define. Since the DMOS values tend to saturate, strict objective limits are useless to draw. Also the objective measures are not capable to evaluate different types of codecs. Nevertheless, the current numerical measures work in some reasonable sense and they are much more practical in the design of quantizers than the listening tests are. However, advanced numerical quantization measures are needed.

It seems that the subjective quality of the LSF quantization in IS-641 codec has a saddle point, and after that point the perceptual quality saturates. This may results from the fact that the worst parts of the codec dominate the overall performance and there is no need to improve the quality of the good parts. With the clean speech in IS-641, the 20-bit quantizer has this saddle point, since it has the same perceptual quality than higher bit rate quantizers. Consequently this gives a doubt that the present transparency limits of SD in literature might be too strict to the IS-641 type of speech codec.

From the three- and the four-excitation-pulse codecs it can be stated that a proper allocation of bits between quantization of spectra and excitation is essential in low bit rate codecs. For example, the three-pulse codec, with 26-bit LSF quantizer, uses 8 bits less in the frame than the four-pulse codec, with 14-bit LSF quantization. That makes the bit rate to decrease 400 bits per second, while the perceptual quality of speech remains the same.

The voiced and unvoiced partition of the frames shows that the subjective quality criterion depends on the speech sounds. The accuracy and tolerance of hearing regarding the quantization error of LP spectra is not well known, but even the basic division to voiced and

unvoiced sounds seems to be effective. The LSF quantization of the unvoiced sound spectra can be loosened, while the voiced sound has to be coded accurately. According to the results, the limit proposed in [53] of unvoiced frame quantization seems to be reasonable. However, in these tests, the limits of voiced frame quantization appear to have a needless margin.

6 Conclusions

The trend for coding digital speech signals at low bit rates is continually expanding. Linear predictive coders are most commonly used because they can efficiently represent speech signal at low bit rates. For every frame of speech, the spectrum of signal is modeled using linear predictive analysis, and the filter coefficients are then encoded and transmitted. Line spectral frequency representation is typically used for the quantization of the filter parameters. In this thesis moving average split vector quantization of these spectral parameters has been studied.

6.1 Contribution of the Thesis

This thesis is a study on the predictive quantization of spectral parameters. Several moving average predictors were evaluated using both objective and subjective measures. The main results of this thesis are, firstly, a quantizer using an inter-split predictor can gain nearly one bit compared to the conventional linear predictive quantizers and secondly, the performance of the quantizer of the IS-641 codec at 26 bits per frame can be achieved with a quantizer using an inter-split predictor at 22 bits per frame.

The training algorithm was evaluated and the training tools were coded. The quantizer structures and the algorithm were presented in Chapter 3. Several hundred quantizers were trained and evaluated to adjust the performance of the quantizer structures. The objective results of these quantizers were shown in Chapter 4. The results show that the two-split

quantizer using the inter-split predictor of order two is the most effective in the context of spectral distortion. In addition two-split quantizers outperform three-split quantizers, on average by one bit.

Finally two subjective listening tests were carried out. The results point out that perceptually transparent quality could be achieved with IS-641 using even a 20-bit predictive quantizer. This brings a doubt that the objective limits of transparent quality might be too strict for the IS-641 type of codec. However, the tests studied only a noiseless environment and more exhaustive tests should be made in background noise conditions. In addition the results indicate that objective and subjective measures correspond to each other, but accurate limits of the perceptually sufficient quantization level of spectral parameters are difficult to define. Furthermore, the quantization of spectral parameters in unvoiced and voiced frames was studied. It was proven that the quantization of voiced frames affects most the perceptual quality, and unvoiced frames can be loosely quantized.

The work in this thesis was done by the author, excluding some mathematical help on presenting the prediction parameter estimation in Section 3.3. In addition the generalized Lloyd training algorithm in Figure 3.1, and the training and validation databases were given. However, the training algorithm for predictive quantizers, presented in Figure 3.2 was coded by the author. In addition the training and validation of the quantizers were done by author. The author also constructed and administered the subjective tests. Nevertheless the instructor helped and contributed to several issues along the work.

6.2 Future work

The study in this thesis arises several questions. Some of them were already answered and some of them not. In addition the outline of the thesis left some issues out of the study. In the following paragraphs, the four most important issues suitable for further study are presented.

The thesis studied the quantizers in errorless channels. Thus studies in transmission error conditions were not included. The robustness of transmission errors was discussed in the context of limiting the error propagation of predictive quantizers. It was concluded that a suitable upper limit for the order of the quantizer would be two. However, proper index assignment of the codebook would increase the overall robustness. In addition, the channel optimized coding is a more powerful tool for the field of transmission errors [54].

Although the spectral distortion measure is the primary choice for evaluating the performance of the quantizers, designing a quantizer that directly minimizes the overall spectral distortion is difficult due the complexity of the measure. In this thesis quantizers are designed with minimizing the Euclidean distance of original and quantized LSF

parameters. The mismatch of these measures weakens the performance of the quantizer. To overcome this problem, better sub-optimal measures could be used. Examples of these methods are found in [55].

In Chapter 5, the voiced and the unvoiced parts of LP spectrum were studied. The results were interesting and would certainly be worth further studies especially in low bit rate coding. The results showed that even using as simple a division as a division in unvoiced and voiced sounds, some parts of the spectrum can be quantized considerably loosely. Some articles have been published in this field [53, 56] that would be good references for advanced study.

Eventually predictive structures develop all the time. Lately the non-linear predictors have shown to be efficient [43, 44]. In addition, several other structures have been presented that certainly would be worth study [57].

References

[1] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: MacMillan, 1993.

[2] N. Jayant and P. Noll, *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Englewood Cliffs, New Jersey: Prentice-Hall, 1984.

[3] R. Salami, C. Laflamme, J. P. Adoul, and D. Massaloux, "A toll quality 8 kb/s speech codec for the personal communications system (PCS)", *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 808–816, August 1994.

[4] R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, and Y. Shoham, "Description of the proposed ITU-T 8-kb/s speech coding standard", in Proc. *IEEE Workshop on Speech Coding for Telecom.*, (Annapolis, MD), pp. 3–4, September 1995.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. Of the IEEE"*, vol. 63, no. 4, pp. 561–580, 1975.

[6] L. Ljung, *System Identification: Theory for the User*, Englewood Cliffs, NJ: Prentice-Hall, pp. 278–280, 1987.

[21] Nippon Telephone and Telegraph Corporation, "Multi-Lingual Speech Database for Telephonometry 1994", Tokyo, Japan, 1994.

[7] T. Honkanen, J. Vainio, K. Järvinen, P. Haavisto, R. Salami, C. Laflamme, and J.-P. Adoul, "Enhanced Full Rate Speech Codec for IS-136 Digital Cellural System", in Proc. *IEEE-International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol. 2, pp. 731–734, (Munich, Germany), 1997.

[8] TIA/EIA/IS-641. TDMA Cellular/PCS - Radio Interface - Enhanced Full-Rate Speech Codec. Telecommunications Industry Association / Electronic Industry Association / Interim Standard 641. Global Engineering Documents, 1996.

[9] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame", *IEEE Transactions On Speech And Audio Processing*, vol. 1, no. 1, 1993.

[10] Y. Tohkura and F. Itakura, “Spectral smoothing technique in PARCOR speech analysis-synthesis,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 6, pp. 587–596, 1978.

[11] F. Itakura, “Line spectrum representation of linear prediction coefficients of speech signal”, *Journal Acoustical Society America*, vol. 57, p. 535, 1975. (abstract).

[12] A. H. Gray, Jr., R. M. Gray, and J. D. Markel, “Comparison of optimal quantizations of reflection coefficients”, *IEEE Transactions on Acousics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 9–23, 1977.

[13] F. K. Soong and B. H. Juang, “Line spectrum pair (LSP) and speech data compression.” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (*San Diego), CA, pp.1.10.1–1.10.4, Mar. 1984.

[14] F. Itakura and N. Sugamura, LSP Speech Synthesizer, Speech Group, *Acuost. Soc. of Japan, Tech. Rept*. 5, 1979.

[15] G. S. Kang and L. J. Fransen, “Application of line-spectrum pairs to low-bit-rate speech encoders”, in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Prosessing (ICASSP)*, (Tampa, FL), pp. 244–247, 1985.

[16] K. K. Paliwal and W. B. Kleijn, “Quantization of LPC Parameters”, in *Speech Coding and Synthesis* (Kleijn and Paliwal eds.), pp. 433–466, Elsevier Science, 1995.

[17] F. Soong and B.-H. Juang, “Optimal quantization of LSP parameters”, *IEEE Transactions on Speech and Audio Processing*, vol. 1, pp. 15–24, Jan. 1993.

[18] P. Kabal and R. P. Ramachandran, “The computation of line spectral frequencies using Chebyshev polynomials,” Proc. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1419–1426, Dec. 1986.

[19] ITU-T Recommendation P.48, “Specification for an intermediate reference system, Volume V of Blue Book”, pp 81–86. *ITU, Geneva*, Feb. 1996.

[20] ITU-T Recommendation P.830, “Subjective performance assessment of Telephone Band and Wideband Digital Codecs”. *ITU, Geneva*, Feb. 1996.

[22] DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus, Department of Commerce, National Institute of Standards and Technology, NTIS, Springfield, Virginia, October 1990.

[23] J. S. Collura, A. McCree, and T. E. Tremain, “Perceptually based distortion measurements for spectrum quantization,” in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Annapolis, MD), pp. 49–50, 1995.

[24] A. H. Gray, Jr. and J. D. Markel, “Distance measures for speech prosessing,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, pp. 380–391, Oct. 1976.

[25] R. Laroia, N. Phamdo, and N. Farvardin, “Robust and effecient quantization of speech LSP parameters using structured vector quantizars”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 641–644, 1991.

[26] F. Tzeng, “Analysis-By-Synthesis Linear Predictive Speech Coding at 2.4 kbit/s,” in Proc. *IEEE Global Telecommunications Conference (Globecom)*, pp. 1253–1257, 1989.

[27] D. Chang, Y. Cho, S. Ann, “Effecient quantization of LSF parameters using conditional splitting scheme”, Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Detroit, MI) , pp. 736–739, May 1995.

[28] E. Paksoy, W.-Y. Chan, and A. Gersho, “Vector quantization of speech LSF parameters with generalized product codes”, in Proc. *IEEE International Conference on Spoken Language Processing*, (Banff, Canada), pp. 33–36, 1992.

[29] J. R. Deller, Jr., J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. New York: MacMillan, 1993.

[30] A. Gersho and V. Cuperman, “Vector quantization: A pattern matching technique for speech coding”, *IEEE Commun. Mag*., vol. 21, pp. 15–21, 1983.

[31] R. M. Gray, “Vector quantization”, *IEEE Acoustics, Speech, and Signal Processing Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[32] J. Makhoul, S. Roucos, and H. Gish, “Vector quantization in speech coding”, *Proc. IEEE*, vol. 73, no. 11, pp. 1551–1588, 1985.

[33] C. E. Shannon, “A mathematical theory of communication”, *Bell Systems Technical Journal*, pp. 27:379–423, 623–656, 1948.

[34] C. E. Shannon, “Coding theorems for a discrete source with a fidelity criterion”, in *Proc. IRE National Convention Rec*., Part 4, pp. 142–163, 1959.

[35] N. Farvardin, and R. Laroia, “Efficient encoding of speech LSP parameters using the discrete cosine transform”, in Proc of *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Glasgow), pp. 168–171, May 1989.

[36] H. Ohmuro, T. Moriya, K. Mano, and S. Miki, “Coding of LSP parameters using interfame moving averave prediction and multi-stage vector quantization”, in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, (Sainte-Adèle), Canada, pp. 63–64, Oct. 1993.

[37] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer Academic Publishers, 1992.

[38] S. P. Lloyd, “Least squares quantization in PCM”, *IEEE Transactions on Information Theory*, vol. 28, pp. 129–137, March 1982.

[39] J. H. Y. Loo, “Intraframe and Interframe Coding of Speech Spectral Parameters”, Master’s thesis, McGill University, Montreal, Canada, 1996.

[40] Y. Linde, A. Buzo and R. M. Gray, “An algorithm for vector quantizer design”, *IEEE Transactions on Communications*, vol. 28, no. 1, pp. 84–95, 1980.

[41] I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang, “A new initialization technique of Generalized Lloyd Algorithm”, *IEEE Signal Processing Letters*, vol. 1, pp. 144–146, Oct. 1994.

[42] R. Hagen, “Robust LPC spectrum quantization - Vector quantization by a linear mapping of a block code”, Doctoral Thesis, Chalmers University of Technology, Sweden, Feb. 1995.

[43] J. H. Y. Loo and W.-Y. Chan, “Nonlinear predictive vector quantization of speech spectral parameters”, in Proc. *IEEE Workshop on Speech Coding for Telecommunications*, (Annapolis, MD), pp.51–52, Sep. 1995.

[44] J. H. Y. Loo, W.-Y. Chan, and P. Kabal, “Classified nonlinear predictive vector quantization of speech parameters”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Atlanta, GA), pp. II-761–II-764, May 1996.

[45] B. Townshend, “Nonlinear prediction of speech”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, (Toronto, Canada), pp. 425–428, May 1991.

[46] J. Thyssen, H. Nielsen, and S. D. Hansen, “Non-linear short-time prediction in speech coding”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP)*, (Adelaide), pp. I-185–I-188, April 1994.

[47] J. Thyssen, H. Nielsen, and S. D. Hansen, “Quantization of non-linear predictors in speech coding”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP)*, (Detroit), pp. 265–268, May 1995.

[48] L. Wu, M. Niranjan, and F. Fallside, “Fully vector-quantized neural network-based code-excited nonlinear predictive speech coding”, *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 482–489, Oct. 1994.

[49] ITU-T. Recommendation P.80, Telephone transmission quality subjective opinion tests. Annex D - Degradation Category Rating. ITU, Helsinki, March 1994.

[50] B. H. Juang, D. Y. Wong, and A. H. Gray, Jr., “Distortion performance of vector quantization for LPC voice coding”, *IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP)*, vol. 30, no. 2, pp. 294–304, 1982.

[51] S. Wang, and A. Gersho, “Phonetically-based vector excitaton coding of speech at 3.6 kbits/s”, in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP)*, (Glasgow), pp. I-369–372, 1989.

[52] S. Wang, and A. Gersho, “Improved phonetically-segmented vector excitation coding at 3.4 kbits/s,” in Proc. *IEEE International Conference on Acoustics, Speech, and Signal Proccessing (ICASSP)*, (San Francisco, CA), pp. I-349–352, 1992.

[53] R. Hagen, E. Paksoy and A. Gersho, “Voicing-Specific LPC Quantization for Variable Rate Speech Coding”, *On Robust LPC-spectrum Coding and Vector Quantization*, Doctoral Thesis, Chalmers University of Technology, Göteborg, Sweden 1995.

[54] J. Lindén, “Channel optimized predictive vector quantization”, *Interframe Quantizatin for Noisy Channels*, Doctoral Thesis, Chalmers University of Technology, Göteborg, Sweden 1998. Paper has been submitted for publication to *IERE Transactions on Speech and Audio Processing*, May 1998.

[55] W. R. Gardner, and B. D. Rao, “Theoretical analysis of the high-rate vector quantization of LPC parameters”, *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 367–281. Sept. 1995.

[56] E. Paksoy, K. Srinivasan, and A. Gersho, “Variable bit-rate CELP coding of speech with phonetic classification,” *European Transactions on Telecommunications*, vol. 5, pp. 591–601, Sep.-Oct. 1994.

[57] T. Eriksson, J. Lindén, and J. Skoglund, “A safety-net approach for improved explotation of speech correlations,” in *Signal Processing VII: Theories and Applications* (M. Holt, C. Cowan, P. Grant, and W. Sandham, eds.), pp. 4–7, 1994.

Appendices

Appendix I Detailed results of the first listening test

Degradation Category Rating of LSF Quantizers
27.6.-8.7.-97 - Results

Author: Markus Vaalgamaa
17.9.-97 Nokia Research Center

Candidates:
The test is based on the IS-641 speech codec where the LSF quantizer is changed. The quality of the LSF quantizers of 14, 17, 20, 23 and 26 bits per frame compared to the original 26-bit LSF quantizer of IS-641 and to unquantized LSF are evaluated. All quantizers have structure 1-DP-3-SVQ. 14-26 bits candidates are trained with both flat and modified IRS filtered speech, where as IS-641 is trained only with modified IRS filtered speech.

Speech material:
The input speech material consisted of six British-English speech samples (three males and three females) from NTT's Multi-Lingual Speech Database for Telephonemetry 1994 speech database. Following to the ITUT-T Recommendation speech samples, each containing one sentence, are modified IRS filtered.

Number of listeners: 37 from 40.

The results:

DMOS-values, Spectral Distortion of LSF quantizers and segmental Signal/Noise Ration are presented in following tables. In DMOS tables a column 'lower' is a 95 % lower reliability limit, and a column 'upper' is a 95 % upper reliability limit.

Voice activity detection is used, meaning that silent frames are removed from the objective results.

Number of frames:

F1	F2	F3	M1	M2	M3	total
157	169	177	142	135	122	902

Samples voice activities in %:

F1	F2	F3	M1	M2	M3	total
67	67	72	67	67	61	67

Listening results [DMOS]

cand.	F1	F2	F3	M1	M2	M3	total	lower	upper
14	3.08	2.89	2.92	3.65	2.78	3.86	3.20	3.09	3.31
17	2.97	2.89	3.14	3.86	3.08	4.14	3.35	3.23	3.46
20	3.08	3.11	3.22	3.84	3.27	4.22	3.45	3.34	3.57
23	3.22	3.05	3.30	3.89	3.08	4.38	3.49	3.37	3.60
26	3.43	3.30	3.24	3.89	3.27	4.19	3.55	3.45	3.66
IS	3.32	3.08	3.41	3.97	3.32	4.05	3.53	3.42	3.63
NQ	3.03	3.27	3.30	4.08	3.24	4.22	3.52	3.41	3.64

MNRU's [DMOS]:

cand.	F1	F2	F3	M1	M2	M3	total	lower	upper
5	1.24	1.14	1.05	1.30	1.08	1.30	1.18	1.13	1.24
15	2.14	2.00	1.76	2.95	1.95	2.19	2.16	2.06	2.26
30	4.51	4.24	4.35	4.65	4.51	4.59	4.48	4.39	4.56
dir	4.86	4.78	4.81	4.76	4.92	5.00	4.86	4.80	4.91

Spectral Distortion

Average spectral distortion [dB]:

cand.	F1	F2	F3	M1	M2	M3	total
14	2.13	2.10	2.09	2.22	2.24	2.18	2.16
17	1.73	1.68	1.73	1.86	1.85	1.77	1.77
20	1.43	1.37	1.42	1.50	1.45	1.48	1.44
23	1.26	1.15	1.13	1.18	1.24	1.25	1.20
26	1.03	0.97	0.94	1.04	1.04	1.02	1.01
IS	1.07	1.08	1.10	1.07	1.11	1.14	1.09
NQ	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Percentage of outliers having SD over 2 dB in %:

cand.	F1	F2	F3	M1	M2	M3	total
14	53.33	58.41	48.03	60.00	62.22	60.81	57.13
17	20.00	21.24	24.41	31.58	31.11	27.03	25.89
20	7.62	6.19	8.66	14.74	14.44	12.16	10.64
23	3.81	0.00	3.94	4.21	4.44	4.05	3.41
26	2.86	0.88	1.57	1.05	0.00	1.35	1.29
IS	0.95	1.77	2.36	1.05	3.33	4.05	2.25
NQ	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Segmental signal-to-noise ratio

SegSNR (from segsnr-matlab function, made by S.Pietilä):

cand.	F1	F2	F3	M1	M2	M3	total
14	9.83	11.10	13.80	8.32	10.62	10.62	10.72
17	10.19	11.37	14.19	8.61	10.72	11.11	11.03
20	10.43	11.52	14.31	8.64	11.09	11.24	11.21
23	10.48	11.65	14.62	8.93	11.28	11.52	11.41
26	10.51	11.69	14.58	9.27	11.36	11.53	11.49
IS	10.48	11.81	14.67	8.93	11.31	11.73	11.49
NQ	10.67	11.95	14.94	9.18	11.68	11.95	11.73

Histogram of votes for each condition

cand.	DMOS-value				
	1	2	3	4	5
14	4	47	102	71	16
17	2	38	98	80	22
20	1	31	88	96	24
23	0	29	96	86	29
26	1	23	90	96	30
IS	0	22	95	101	22
NQ	3	23	94	90	30

Appendix II Detailed results of the second listening test

Relation of voiced and unvoiced frames LSF quantization
and effect of excitation signal quality
25.8.-3.9.-97 - Results

Author: Markus Vaalgamaa
18.9.-97 Nokia Research Center

Candidates:
The test is based on the IS-641 speech codec in which the LSF quantizer is changed. The qualities of codecs using 1-DP-3-SVQs of bit rates 14, 17, 20, 23 and 26 bits per frame are evaluated in four cases.

In the first case a codec using 3 excitation pulses (12 bits per subframe) is evaluated and in the second case a codec using 4 excitation pulses (17 bits per subframe), is evaluated. The last two cases are based on 4 excitation pulse codec. Either unvoiced or voiced parts of spectrum are quantized.

Speech material:
The input speech material consisted of six British-English speech samples (three males and three females) from NTT's Multi-Lingual Speech Database for Telephonemetry 1994 speech database. Following to the ITUT-T Recommendation speech samples, each containing one sentence, are modified IRS filtered and processed. The evaluation is made using high-quality headphones.

Number of listeners: 24 from 26.

The results:

DMOS-values, Spectral distortion of LSF quantizers and segmental Signal/Noise Ration are presented in following tables. In DMOS tables a column 'lower' is a 95 % lower reliability limit, and a column 'upper' is a 95 % upper reliability limit.

Voice activity detection is used, meaning that silent frames are removed from the objective results.

Number of frames:

F1	F2	F3	M1	M2	M3	total
164	169	177	142	135	130	917

Samples voice activities in %:

F1	F2	F3	M1	M2	M3	total
65	67	72	67	67	65	67

LSF quantizers in 3 and 4 excitation pulse codec

Vad flag is used, meaning that silent frames are removed when calculating objective results.

Listening test results [DMOS]:

cand.	F1	F2	F3	M1	M2	M3	total	lower	upper
3_14	2.88	2.29	2.62	3.46	2.42	2.67	2.72	2.60	2.84
3_17	3.00	2.42	2.75	3.33	2.62	2.88	2.83	2.71	2.96
3_20	3.00	2.46	2.88	3.71	2.83	3.00	2.98	2.84	3.12
3_23	3.00	2.58	2.58	3.88	2.46	3.50	3.00	2.86	3.14
3_26	2.79	2.83	2.96	4.08	2.79	3.33	3.13	2.99	3.27
4_14	3.29	2.79	3.00	3.62	2.50	3.67	3.15	3.01	3.29
4_17	3.62	2.79	3.12	4.29	2.96	3.58	3.40	3.25	3.54
4_20	3.50	3.17	3.17	4.29	2.88	4.04	3.51	3.36	3.65
4_23	3.67	2.88	3.12	4.04	2.88	4.21	3.47	3.32	3.61
4_26	3.62	3.17	3.29	4.04	3.21	3.88	3.53	3.40	3.67

MNRU samples of noise level (dB) [DMOS]:

cand.	F1	F2	F3	M1	M2	M3	total	lower	upper
10	1.12	1.12	1.08	1.42	1.08	1.29	1.19	1.12	1.25
20	2.12	2.17	1.96	3.67	2.12	2.71	2.46	2.30	2.61
30	4.08	4.04	3.96	4.71	4.25	4.67	4.28	4.16	4.41
dir	5.00	4.71	4.92	4.88	4.88	4.96	4.89	4.83	4.95

Average spectral distortion [dB], same for both 3 and 4 pulse codecs:

bits	F1	F2	F3	M1	M2	M3	total
14	2.29	2.10	2.09	2.22	2.24	2.13	2.18
17	1.84	1.68	1.73	1.86	1.85	1.77	1.79
20	1.60	1.37	1.42	1.50	1.45	1.39	1.46
23	1.31	1.15	1.13	1.18	1.24	1.17	1.20
26	1.06	0.97	0.94	1.04	1.04	1.00	1.01

Percentage of outliers having SD over 2 dB in %:

bits	F1	F2	F3	M1	M2	M3	total
14	62.62	58.41	48.03	60.00	62.22	58.33	58.27
17	26.17	21.24	24.41	31.58	31.11	28.57	27.18
20	13.08	6.19	8.66	14.74	14.44	7.14	10.71
23	5.61	0.00	3.94	4.21	4.44	1.19	3.23
26	3.74	0.88	1.57	1.05	0.00	1.19	1.41

Segmental signal-noise-ration (from segsnr-matlab function, made by S.Pietilä):

cand.	F1	F2	F3	M1	M2	M3	total
3_14	8.99	9.74	12.65	6.95	9.15	8.84	9.39
3_17	9.24	10.04	12.72	7.24	9.41	9.23	9.65
3_20	9.40	10.33	13.16	7.49	9.64	9.59	9.93
3_23	9.63	10.44	13.32	7.57	9.93	9.66	10.09
3_26	9.63	10.54	13.42	7.75	9.88	9.74	10.16
4_14	10.20	11.10	13.80	8.32	10.62	10.26	10.72
4_17	10.63	11.37	14.19	8.61	10.72	10.58	11.02
4_20	10.79	11.52	14.31	8.64	11.09	10.89	11.21
4_23	10.98	11.65	14.62	8.93	11.28	11.02	11.41
4_26	11.04	11.69	14.58	9.27	11.36	11.02	11.50

LSF:s in UNVOICED frames are quantized

Vad flag (voice activity) and vuv (voiced/unvoidec) flag are used.

Samples voice activities in %:

F1	F2	F3	M1	M2	M3	total
65	67	72	67	67	65	67

Unvoiced frames from voice active frames %:

F1	F2	F3	M1	M2	M3	total
32	26	11	38	23	30	27

Listening results in DMOS:

bits	F1	F2	F3	M1	M2	M3	total	lower	upper
14	3.71	3.08	3.21	4.21	3.17	3.71	3.51	3.38	3.64
17	3.67	2.83	3.17	4.08	3.25	3.67	3.44	3.31	3.58
20	3.62	3.12	3.38	4.42	2.83	3.92	3.55	3.41	3.69
23	3.83	2.79	3.42	4.42	3.08	4.00	3.59	3.44	3.74

```
 26    3.54     2.96     3.29     4.17     3.29     4.00     3.54    3.40   3.68
Average spectral distortion [dB], calculated from unvoiced frames:
 bits   F1       F2       F3       M1       M2       M3      total
-----------------------------------------------------------------
 14    2.39     2.19     2.26     2.08     2.35     2.13     2.24
 17    1.93     1.50     1.51     1.32     1.68     1.62     1.59
 20    1.69     1.43     1.55     1.46     1.53     1.45     1.52
 23    1.35     1.26     1.37     1.07     1.34     1.19     1.26
 26    1.14     1.05     1.00     1.00     1.12     0.96     1.04

Percentage of outliers having SD over 2 dB in %, calculated from unvoiced frames:
 bits   F1       F2       F3       M1       M2       M3      total
-----------------------------------------------------------------
 14    73.53    65.52    57.14    55.56    66.67    56.00    62.40
 17    44.12    27.59    28.57    16.67    33.33    24.00    29.05
 20    17.65    10.34    7.14     11.11    14.29    8.00     11.42
 23    5.88     0.00     14.29    0.00     9.52     4.00     5.62
 26    5.88     3.45     7.14     0.00     0.00     4.00     3.41

Segmental signal/noise ration (from segsnr-matlab function),
calculated from active frames:
 cand.  F1       F2       F3       M1       M2       M3      total
-----------------------------------------------------------------
 14    11.24    11.92    14.87    9.14     11.50    11.01    11.61
 17    11.22    11.91    14.81    9.12     11.41    11.07    11.59
 20    11.18    11.94    14.87    9.28     11.48    11.12    11.65
 23    11.19    11.84    14.78    9.34     11.41    11.21    11.63
 26    11.40    11.95    14.99    9.29     11.34    11.30    11.71
```

LSF:s in VOICED frames are quantized

Vad flag (voice activity) and vuv (voiced/unvoidec) flag are used.

```
Samples voice activities in %:
        F1       F2       F3       M1       M2       M3      total
-----------------------------------------------------------------
        65       67       72       67       67       65       67

Voiced frames from voice active frames %:
        F1       F2       F3       M1       M2       M3      total
-----------------------------------------------------------------
        68       74       89       62       77       70       73
```

Listening results in DMOS:

bits	F1	F2	F3	M1	M2	M3	total	lower	upper
14	3.38	2.92	2.96	3.75	3.00	3.79	3.30	3.17	3.43
17	3.75	3.04	2.75	4.21	3.08	3.92	3.46	3.32	3.60
20	3.88	3.21	3.04	4.12	2.96	3.71	3.49	3.35	3.62
23	3.92	2.96	3.46	4.38	3.38	3.71	3.63	3.49	3.77
26	3.79	2.88	3.29	4.00	3.21	3.96	3.52	3.39	3.65

Average spectral distortion [dB], calculated from voiced frames:

bits	F1	F2	F3	M1	M2	M3	total
14	2.24	2.06	2.07	2.30	2.21	2.13	2.17
17	1.84	1.68	1.66	1.78	1.85	1.72	1.76
20	1.56	1.34	1.41	1.52	1.43	1.37	1.44
23	1.29	1.12	1.10	1.24	1.21	1.17	1.19
26	1.02	0.94	0.94	1.07	1.01	1.01	1.00

Percentage of outliers having SD over 2 dB in %, calculated from voiced frames:

bits	F1	F2	F3	M1	M2	M3	total
14	57.53	55.95	46.90	62.71	60.87	59.32	57.22
17	32.88	21.43	12.39	25.42	31.88	18.64	23.77
20	10.96	4.76	8.85	16.95	14.49	6.78	10.47
23	5.48	0.00	2.65	6.78	2.90	0.00	2.97
26	2.74	0.00	0.88	1.69	0.00	0.00	0.89

Segmental signal/noise ration (from segsnr-matlab function),
calculated from active frames:

cand.	F1	F2	F3	M1	M2	M3	total
14	10.57	11.12	14.12	8.50	10.69	10.67	10.95
17	10.72	11.42	14.15	8.82	10.83	10.87	11.14
20	11.00	11.67	14.51	8.87	11.09	10.75	11.31
23	11.02	11.64	14.50	9.02	11.28	10.96	11.40
26	11.01	11.76	14.82	9.07	11.32	10.97	11.49

Histogram of votes for each condition

cand.	DMOS-value				
	1	2	3	4	5
4_14	2	31	62	42	7
4_17	0	20	63	45	16
4_20	1	15	58	50	20
4_23	1	13	67	44	19
4_26	1	13	53	62	15
3_14	4	50	74	14	2
3_17	3	44	76	16	5
3_20	2	38	72	25	7
3_23	2	37	71	27	7
3_26	2	30	68	35	9
14(u)	0	14	55	62	13
17(u)	1	15	61	53	14
20(u)	0	13	61	48	22
23(u)	0	15	55	48	26
26(u)	0	15	54	57	18
14(v)	0	20	72	41	11
17(v)	1	17	57	53	16
20(v)	0	15	59	55	15
23(v)	0	13	49	60	22
26(v)	0	13	60	54	17