The development of speech coding and the first standard coder for public mobile telephony

The Development of Speech Coding and the First Standard Coder for Public Mobile Telephony

R.J. Sluijter

The work described in this thesis has been carried out
at the PHILIPS RESEARCH LABORATORIES Eindhoven,
the Netherlands, as part of the Philips Research Programme.

CIP-gegevens Koninklijke Bibliotheek, Den Haag
Sluijter, R.J.
The Development of Speech Coding and the First Standard Coder
for Public Mobile Telephony
Proefschrift Technische Universiteit Eindhoven,-Met lit. opg.,
-Met samenvatting in het Nederlands.
ISBN 90-74445-00-4
Trefw.: speech coding, speech codec, voice communication, history,
cellular radio, mobile telephony, GSM standard

Cover design: Bertina Senders

The Development of Speech Coding and the First Standard Coder for Public Mobile Telephony

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof.dr.ir. C.J. van Duijn,
voor een commissie aangewezen door het college voor Promoties
in het openbaar te verdedigen
op maandag 10 oktober 2005 om 16.00 uur

door

Robert Johannes Sluijter

geboren te Nijmegen

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. J.W.M. Bergmans
en
prof.Dr.-Ing. P. Vary

To: *my wife José*
our son Jeroen and his wife Ingrid,
and their children Floor and Luuk
our daughter Babette
our son Maarten
our future descendants

Preface

Summary

This thesis describes in its core chapter (Chapter 4) the original algorithmic and design features of the first coder for public mobile telephony, the GSM full-rate speech coder, as standardized in 1988. It has never been described in so much detail as presented here. The coder is put in a historical perspective by two preceding chapters on the history of speech production models and the development of speech coding techniques until the mid 1980s, respectively. In the epilogue a brief review is given of later developments in speech coding.

The introductory Chapter 1 starts with some preliminaries. It is defined what speech coding is and the reader is introduced to speech coding standards and the standardization institutes which set them. Then, the attributes of a speech coder playing a role in standardization are explained. Subsequently, several applications of speech coders - including mobile telephony - will be discussed and the state of the art in speech coding will be illustrated on the basis of some worldwide recognized standards.

Chapter 2 starts with a summary of the features of speech signals and their source, the human speech organ. Then, historical models of speech production which form the basis of different kinds of modern speech coders are discussed. Starting with a review of ancient mechanical models, we will arrive at the electrical source-filter model of the 1930s. Subsequently, the acoustic-tube models as they arose in the 1950s and 1960s are discussed. Finally the 1970s are reviewed which brought the discrete-time filter model on the basis of linear prediction. In a unique way the logical sequencing of these models is exposed, and the links are discussed. Whereas the historical models are discussed in a narrative style, the acoustic tube models and the linear prediction tech-

nique as applied to speech, are subject to more mathematical analysis in order to create a sound basis for the treatise of Chapter 4. This trend continues in Chapter 3, whenever instrumental in completing that basis.

In Chapter 3 the reader is taken by the hand on a guided tour through time during which successive speech coding methods pass in review. In an original way special attention is paid to the evolutionary aspect. Specifically, for each newly proposed method it is discussed what it added to the known techniques of the time. After presenting the relevant predecessors starting with Pulse Code Modulation (PCM) and the early vocoders of the 1930s, we will arrive at Residual-Excited Linear Predictive (RELP) coders, Analysis-by-Synthesis systems and Regular-Pulse Excitation in 1984. The latter forms the basis of the GSM full-rate coder.

In Chapter 4, which constitutes the core of this thesis, explicit forms of Multi-Pulse Excited (MPE) and Regular-Pulse Excited (RPE) analysis-by-synthesis coding systems are developed. Starting from current pulse-amplitude computation methods in 1984, which included solving sets of equations (typically of order 10-16) two hundred times a second, several explicit-form designs are considered by which solving sets of equations in real time is avoided. Then, the design of a specific explicit-form RPE coder and an associated efficient architecture are described. The explicit forms and the resulting architectural features have never been published in so much detail as presented here. Implementation of such a codec enabled real-time operation on a state-of-the-art single-chip digital signal processor of the time. This coder, at a bit rate of 13 kbit/s, has been selected as the Full-Rate GSM standard in 1988. Its performance is recapitulated.

Chapter 5 is an epilogue briefly reviewing the major developments in speech coding technology after 1988. Many speech coding standards have been set, for mobile telephony as well as for other applications, since then. The chapter is concluded by an outlook.

Originality Statement

While I was concentrating on RELP coding with down-sampling in the early 1980s, I co-invented RPE together with Ed F. Deprettere and Peter Kroon (doing his Ph.D. work in which I was involved) of the Technical University of Delft. RPE is also based on down-sampling, but in an analysis-by-synthesis framework. The basic (Philips) patent on RPE

carries those two names and my name as inventors.

Subsequently, I invented several explicit forms - which are described in the same patent - and I designed a corresponding RPE coding architecture, as described in Chapter 4, which enabled the real-time implementation of the complete codec on a single-chip digital signal processor of the time.

I was also involved in the design of such a digital signal processor together with Peter Vary and Karl Hellwig of Philips Kommunikations Industrie (PKI) Nuremberg, Peter Anders of Philips Semiconductors Hamburg, and Jef van Meerbergen, Frank Welten and Frans van Wijk of Philips Research Eindhoven. This made me very well aware of the possibilities and limitations of these processors.

Based on this knowledge, I designed finite-word-length algorithms for a specific explicit-form RPE coder which could readily be mapped onto any state-of-the-art signal processor of the time.

I transferred the coder software and the implementation know-how to Peter Vary, who was also involved in the creation process of digital mobile telephony, and he proposed the coder to the CEPT[1] as a German candidate for GSM.

Several modifications, which do not belong to my original work, have been introduced to this coder during the standardization process. It concerns the contribution of C. Galand and M. Rosso of IBM France who integrated a pitch predictor into the coder architecture, as described in Section 4.4.1, though the issue of adding a pitch predictor was already covered by the basic RPE patent. It also concerns some design details described in Section 4.5. The basic coder architecture and the rest of the design details do belong to my original work.

This coder became the first speech coder ever standardized for digital public mobile telephony as part of the Full-Rate GSM standard in 1988. It was selected out of six candidates from France, Germany, Italy, Norway, Sweden and the United Kingdom. It has been applied in several billions of mobile telephone handsets, as well as in the infrastructure of the networks all over the world, since then. And still it is today.

[1] GSM stood for Groupe Spéciale Mobile and it was part of the CEPT (Conference of European Posts and Telecommunications administrations). In 1988, it has been merged in a new organization with the name ETSI (European Telecommunications Standards Institute). Later, during the growing adoption of the system on a global scale, GSM was renamed as Global System for Mobile.

Acknowledgements

I am especially grateful to Piet van Gerwen, previously of Philips Research Laboratories, who taught me to think in a scientific way, who introduced me into the field of speech coding and who guided me for many years like a father. I am also grateful to my first group leader Frank de Jager for adopting me in his research group and especially to the vision and support of his successor Leo Zegers, who directed me and others to the field of mobile telephony, at a time that the importance of this field was still hard to defend. Freek Valster and later other Philips Research managers, including Ben Waumans, Theo Claasen, Hans Brandsma, Hans Peek, Peter van Otterloo, Rick Harwig, Fred Boekhorst and Carel-Jan van Driel, never ceased to support my activities. I thank many previous colleagues, too many to mention without forgetting at least some of them, for many animated, constructive discussions. Special thanks are for Niek Verhoeckx, who raised my accurateness in formulating to a level that even I could appreciate my own writings.

I thank Ed Deprettere and Peter Kroon, previously of the Technical University Delft (TUD), for their contributions to the invention of regular-pulse excitation, and for many fruitful meetings and discussions, initiated and stimulated by Patrick Dewilde (TUD), Theo Claasen and later Hans Peek, both previously of Philips Research.

My special thanks are for Peter Vary, previously of Philips Kommunikations Industrie Nürnberg, now with the Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen, and for his former colleagues who transformed a research activity into a European standard.

I thank Jan Bergmans for inspiring me to write this thesis and his support during the development of it, and I especially appreciate the admission to promotion from the "College voor Promoties", all of the Technical University Eindhoven.

I am grateful to Jan Bergmans, Peter Vary, Armin Kohlrausch, Bastiaan Kleijn, Raymond Veldhuis, Leo Vogten, Bert den Brinker and especially to my colleague and friend since 1965, Ad van den Enden, for reading, criticising and improving draft versions of this thesis.

I am especially grateful to my wife José for her unremitting encouragement and support of all my previous and present professional deployments.

Contents

Glossary

List of Abbreviations

3GPP	Third Generation Partnership Project (GSM based)
3GPP2	Third Generation Partnership Project (non-GSM)
AAC	Advanced Audio Coder
ACELP	Algebraic CELP
ACELP/TCX	ACELP/Transform-Coded eXcitation
ACR	Absolute Category Rating
AD	Analog-Digital
ADM	Adaptive Delta Modulation
ADPCM	Adaptive Differential PCM
AES	Audio Engineering Society
AMDF	Average Magnitude Difference Function
AMR	Adaptive Multi-Rate
ANSI	American National Standards Institute
APCM	Adaptive Pulse Code Modulation
ASIC	Application Specific Integrated Circuit
BPF	Band-Pass Filter
CCIR	International Consultive Committee for Radio
CCITT	International Consultive Committee for Telegraphy and Telephony
CCR	Comparison Category Rating
CD	Compact Disk
CDMA	Code Division Multiple Access
CELP	Code-Excited Linear Prediction
CEPT	Conference of European Posts and Telecommunications administrations
CMOS	CCR-MOS
CPU	Central Processing Unit

CS-ACELP	Conjugate-Structure ACELP
CVSD	Continuous Variable Slope Delta-modulation
DA	Digital-Analog
DAM	Diagnostic Acceptability Measure
DC	Direct Current
DCDM	Digitally Controlled Delta Modulation
DCME	Digital Circuit Multiplication Equipment
DCR	Degradation Category Rating
DECT	Digital European Cordless Telephone
DEMUX	Demultiplexer
DFT	Discrete-time Fourier Transform (for periodic signals)
DMOS	DCR-MOS
DoD	Department of Defense (USA)
DPCM	Differential Pulse Code Modulation
D*PCM	open-loop DPCM
DRT	Diagnostic Rhyme Test
DSI	Digital Speech Interpolation
DSP	Digital Signal Processor
EFR	Enhanced Full Rate
ETSI	European Telecommunications Standards Institute
EVRC	Enhanced Variable Rate Coder
eX-CELP	extended CELP
FDM	Frequency Division Multiplex
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FM	Frequency Modulation
FR	Full Rate
FS...	Federal Standard ...
GSM	Global System for Mobile
HE-AAC	High-Efficiency AAC
HFR	High-Frequency Regeneration
HR	Half Rate
HVXC	Harmonic Vector Excitation Coder
IC	Integrated Circuit
IEC	International Electrotechnical Commission
IIR	Infinite Impulse Response
IMBE	Improved Multi-Band Excitation
INMARSAT	International Maritime Satellite
INT	Interpolator

I/O	Input/Output
IP	Internet Protocol
IRS	Intermediate Reference System
IS-...	Intermediate Standard ... (TIA)
ISDN	Integrated Services Digital Network
ISO	International Standardization Organization
ITU	International Telecommunication Union
ITU-R	ITU Radiocommunications Sector
ITU-T	ITU Telecommunications Standardization Sector
LD-CELP	Low-Delay CELP
logPCM	logarithmic Pulse Code Modulation
LP	Linear Prediction
LPC	Linear Predictive Coding
LPF	Low-Pass Filter
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
LTP	Long-Term Prediction
MD	Minimum Deviation (quantization method)
MELP	Mixed-Excitation Linear Prediction (vocoder)
MIPS	Million Instructions Per Second
MMSE	Minimum Mean Square Error
MNRU	Modulated Noise Reference Unit
MOPS	Million Operations Per Second
MOS	Mean Opinion Score
MPE	Multi-Pulse Excitation
MPEG	Motion Pictures Expert Group
MSK	Minimum-Shift Keying
MUX	Multipexer
NATO	North Atlantic Trust Organization
NB	Narrow Band
NFC	Noise Feedback Coding
NL	Non-Linear network/circuit
PABX	Private Automatic Branche Exchange
PAM	Pulse Amplitude Modulation
PARCOR	Partial Correlation
PBX	Private Branche Exchange
PC	Personal Computer
PCM	Pulse Code Modulation
PDC	Personal Digital Cellular

pdf	probability density function
PMR	Private Mobile Radio
POTS	Plain Old Telephone Service
PPROC	Parallel Processing (pitch detection method)
PSI-CELP	Pitch Synchronous Innovation CELP
PSTN	Public Switched Telephone Network
Q	Quantizer
Q&C	Quantization and Coding
QCELP	Qualcom CELP
QoS	Quality of Service
RAM	Random Access Memory
RC-active	Resistor Capacitor active (filter)
RCR	Research & Development Center for Radio Systems (Tokyo)
RCELP	Relaxed CELP
RELP	Residual-Excited Linear Prediction
ROM	Read Only Memory
RP	Reference Procedure of LAR quantization
RPE	Regular Pulse Excitation
SB	Sub-Band
SBC	Sub-Band Coder
SIFT	Simplified Inverse Filter Tracking (of pitch)
SNR	Signal to Noise Ratio
SMV	Selectable Mode Vocoder
SRELP	Second-Residual Excited Linear Prediction
SSC	SinuSoidal Coding/Coder
STP	Short Term Prediction
STU	Secure Telephone Unit
TASI	Time Assignment Speech Interpolation
TC	Transform Coding/Coder
TDM	Time Division Multiplex
TDMA	Time Division Multiplex Access
TETRA	Trans-European Trunked Radio
TIA	Telecommunications Industry Association (USA)
UMTS	Universal Mobile Telecommunications System
UNESCO	United Nations Economic Scientific and Cultural Organization
US	Uniform Sensitivity (quantization method)
UV	Ultra Violet

VMR-WB	Variable-Rate Multi-mode Wide-Band (3GPP2)
VoIP	Voice over IP
VQ	Vector Quantiser/Quantisation
VSELP	Vector-Sum Excited Linear Prediction
WB	Wide Band
wMOPS	weighted Million Operations Per Second

List of Symbols

$\mathcal{A}_m$	cross-sectional area of tube section m
$A(z)$	transfer function of the LP inverse filter
$A_m(z)$	z-transform of $\alpha_m[n]$, $A(z)$ of order m
$A(z/\gamma)$	bandwidth expanded $A(z)$ with expansion factor γ
$1/A(z)$	transfer function of the LP synthesis filter
$1/A(z/\gamma)$	bandwidth expanded $1/A(z)$ with expansion factor γ
$AMDF[n]$	average magnitude difference function
a_i	prediction coefficients, a-parameters
$\alpha_m[n]$	impulse response of the LP inverse filter of order m
$\arg\max_x f(x)$	value of x for which $f(x)$ is maximum
$\mathbf{C}$	symmetrical covariance matrix
$\mathcal{C}$	specific capacitance of a transmisssion line
D	decimation factor
$\downarrow D$	downsampling by a factor D
$\uparrow D$	upsampling by a factor D
$\mathcal{D}$	spectral deviation
$\hat{\mathcal{D}}$	spectral deviation upperbound
$\delta(t)$	Dirac impulse
Δx	increment of x
$\mathcal{E}(x)$	expectation of x
$\mathcal{E}(\hat{\mathcal{D}})$	expected or mean spectral deviation upperbound
$e[n]$	error sequence
$\epsilon[n]$	pitch prediction residual signal
η	adiabatic constant of a gass
$\mathcal{F}_x$	exponential spectral flatness of the signal $x[n]$
f	natural frequency (Hertz)
f_0	pitch frequency
f_s	natural sampling frequency
G_P	prediction gain

γ	bandwidth-expansion factor
$H(z)$	transfer function, z-transform of $h[n]$
$h[n]$	impulse response
I^+	amplitude of incident current-wave
I^-	amplitude of reflected current-wave
$\Im(x)$	imaginary part of x
$I(x,\omega)$	Fourier transform of $i(x,t)$
$i(x,t)$	curent of an electric wave at location x and time t
$i^+(t)$	incident current wave
$i^-(t)$	reflected current wave
L	LTP delay in number of samples
$\mathcal{L}$	specific inductance of a transmisssion line
LAR	log area ratio
l	length of a tube section
l_m	length of tube section m
$\lambda[n]$	triangular window in radical truncation
Λ	length of $\lambda[n]$
M	prediction order or highest index tube/filter section
ω	angle frequency (radians per second)
$P(z)$	LTP inverse filter transfer function
$\mathcal{P}(z)$	symmetrical polynomial of extended inverse filter
$1/P(z)$	LTP synthesis filter transfer function
$p(x,t)$	pressure of an acoustic wave at location x and time t
p_k^*	complex conjugate pole of the k^{th} pole p_k
φ	phase
$\dot{\varphi}$	time-derivative of φ
ϕ	RPE grid position
$q[n]$	quantization error
$\mathcal{Q}(z)$	antisymmetrical polynomial of extended inverse filter
$\mathbf{R}$	symmetrical autocorrelation matrix
$\Re(x)$	real part of x
r_m	reflection coefficient of section m
ϱ	density of a gass
$\rho[i]$	autocorrelation coefficient for lag i
$S(z)$	z-transform of $s[n]$
$(\Delta\mathcal{S})_{\max}$	maximum spectral deviation
SNR	signal to noise ratio
SNR_Q	signal to noise ratio of quantizer Q
$s[n]$	discrete-time speech signal

$s(t)$	continuous-time speech signal
$\mathrm{sgn}(x)$	signum function, same as $\mathrm{sign}(x)$
$\mathrm{sign}(x)$	number with sign of x and unity magnitude
$\mathrm{spf}(x)$	spectral flatness of the signal $x[n]$
σ_x^2	variance of x
T	sampling period
t	time
τ	delay
θ	normalized radian frequency ωT
$U(x,\omega)$	Fourier transform of $u(x,t)$
$u(x,t)$	volume velocity of a wave at location x and time t
$u_m^+(t)$	incident volume velocity wave at section m
$u_m^-(t)$	reflected volume velocity wave at section m
$\upsilon(t)$	continuous-time unit step function
$\upsilon[n]$	discrete-time unit step function
$v(x,t)$	voltage of a wave at location x and time t
$W(z)$	weighting filter transfer function
$X(z)$	z-transform of $x[n]$
$X(e^{j\omega T})$	discrete-time Fourier transform (FTD) of $x[n]$
$x[n]$	discrete-time signal
$x(t)$	continuous-time signal
$\overline{x}$	transmission parameter conveying x
$\tilde{x}$	predicted value of x
$x * y$	convolution
Z_0	characteristic impedance

List of Figures

List of Tables

Chapter 1

Introduction

This chapter starts with some preliminaries. It is defined what speech coding is and the reader is introduced to speech coding standards and the standardization institutes which set them. Then, the attributes of a speech coder playing a role in standardization are explained. Subsequently, several applications of speech coders - including mobile telephony - will be discussed and the state of the art in speech coding will be illustrated on the basis of some worldwide recognized standards.

1.1 Some preliminaries

Speech coding could be described as the conversion of an analog speech signal into a digital signal, but such a description better suits the usual definition of an analog to digital (AD) converter. Speech coding means the conversion of a speech signal which already has been digitalized, into another digital signal featuring a lower bit rate than the original signal. Sometimes, it is also referred to as "transcoding" and also as "compression" or "bit rate reduction". Often, "speech coding" does not only mean encoding a speech signal, but rather the complete process including decoding. The words "speech coder" mostly refer in a similar way to the decoder as well. Sometimes, also the word "codec" is used which is a concatenation of the words coder and decoder, and not of the words *en*coder and decoder, as one would expect. The intention of what is meant by a particular designation will always be clear from the context, however.

The environment of a speech codec is often as depicted in Figure 1.1. In this example the situation is like in a GSM telephone. Speech enters

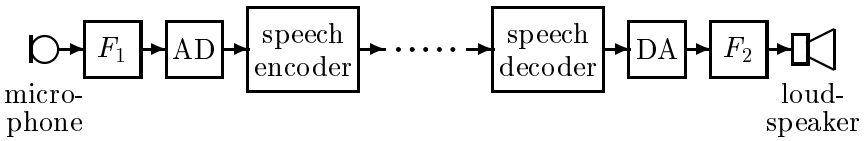

Figure 1.1: System environment of a speech codec.

the microphone and is at first presented to an anti-aliasing filter (F_1). This is the usual preparation to the AD conversion step which consists of sampling and quantization. The anti-aliasing filter limits the bandwidth of the signal to half the sampling frequency which, according to the well-known sampling criterion, is required to enable a faithful reconstruction of the signal. The sampling criterion was already described by H. Nyquist in 1928 for a signal with a finite number of levels [1] and it was generalized in 1948 by Oliver, Pierce and Shannon [2]. In telephony the standard bandwidth of the speech signal is nominally 4 kHz, actually often 300 - 3400 Hz. The sampling frequency amounts to 8 kHz and each sample is quantized into 256 discrete levels and represented by an 8-bit code [3]. The resulting digital signal is called a PCM (from pulse code modulation) signal. This PCM signal is presented to the speech encoder and the compressed signal is transmitted or stored for later usage. In the various applications of speech compression, bit rates may be encountered varying from 64 kbit/s down to only a few hundred bits per second. Decoding results in a PCM signal again which is presented to a digital to analog (DA) converter. After the analog signal has been fed through a so-called reconstruction filter (F_2) it can be presented to the loudspeaker. The reconstruction filter is a low-pass filter again with a passband of (slightly less than) 4 kHz, in the current example. Sometimes, the environment of a speech codec is different, such as for instance in a GSM network. The received signal is decoded as above, but the resulting PCM signal may then directly be presented to the digital infrastructure of the PSTN (public switched telephone network) for transmission to a remote destination.

Many applications of speech coders are in the area of telephony. Telephony is based on a worldwide communications network which is built on international agreements such that inter-operability between the various national networks is guaranteed. These agreements are laid down in standards which are defined by international and regional standardization organizations [4]. International standardization organizations

are e.g. the International Telecommunication Union (ITU), the International Standardization Organization (ISO), the International Electrotechnical Commission (IEC) and the International Maritime Satellite (INMARSAT) organization. Members of these organizations are national governments, scientific organizations and other professional as well as industrial organizations. The ITU is an agency of the worldwide United Nations (Economic, Scientific and Cultural) Organization - UN(ESC)O - residing in Geneva, Switzerland, and all worldwide standardization of telecommunications, including radio communications, is covered by the ITU Telecommunications Standardization Sector (ITU-T). The organization of the ITU and its history is described in an article by Theodor Irmer [5]. Irmer reported that a predecessor of the ITU was founded in 1865 during the early days of telegraphy while telephony did not yet exist at all and "A number of European governments met in Paris to discuss how the upcoming telegraph could be made internationally accessible to the public". Already in 1868, the new organization settled in Bern, Switzerland. As a result of a re-organization of the ITU in 1993, the ITU-T was composed out of the former CCITT (International Consultive Committee for Telegraphy and Telephony) and a part of the former CCIR (International Consultive Committee for Radio). The remaining part was dubbed ITU-R (Radio) and it covers broadcast issues.

For some applications, worldwide standardization is not required and it suffices to work with regional standards which are defined by regional institutes. In Europe, for instance, The European Telecommunications Standards Institute (ETSI) plays an important role in the area of, mainly mobile, telephony and data transmission. In a paper by Gerard Robin [6] the objectives of the ETSI and the interaction with other organizations such as the ITU, the ISO and the IEC, are described. A similar organization is a branch of the ANSI (American National Standards Institute), the Telecommunications Industry Association (TIA). In Japan, standardization for mobile communications is actually accomplished by the Research & Development Center for Radio Systems (RCR) in Tokyo. The RCR is a non-profit organization operating under direction of the Ministry of Posts & Telecommunications.

An overview of what is involved in the standardization of a speech coder is given by Spiros Dimolitsas [7]. It involves the specification of the desired attributes of the speech coder and the specification of the performance evaluation procedures. Finally, verification by third parties of the provisional standard specifications has to guarantee their accuracy

and unambiguousness.

1.2 Attributes of a speech coder

The key attributes which characterize the behaviour of a speech coding system are: the bandwidth of the transferred speech signal, the particular compression method used, the bit rate of the transmitted signal, the speech quality at the decoding side, the transparency of the system, the robustness, the delay, the implementation complexity and the scalability of some of its parameters. These will be all discussed in the following subsections.

1.2.1 Bandwidth

The conventional bandwidth of the speech signal in telephony is 4 kHz. This bandwidth, which is often referred to as narrow-band, is only a nominal indication and in practical systems it is always less than this value. The ITU-T standard P.48 describes an "intermediate reference system" (IRS), which specifies the standard transfer function of narrow-band handsets, for the transmitting as well as for the receiving sides [8]. The lower (3 dB) cutoff frequency of this bandwidth is somewhere between 200 and 300 Hz and the higher cutoff frequency lies between 3400 and 3600 Hz. The recommended sampling frequency for a narrow-band speech signal is 8 kHz, as mentioned before. A "wide band" speech signal, on the contrary, is sampled at 16 kHz. The (3 dB) bandwidth of a standard wide-band speech signal ranges from 50 Hz up to 7 kHz. Wide-band and narrow-band speech coders each form a separate category of coders. Narrow-band speech is the conventional quality when handsets are used, but if the representation of the received speech signal is via a loudspeaker wide-band quality is always preferred.

1.2.2 Coding method

Speech coding methods are often based on models of the human speech production organ and of speech perception (the models of speech production are the subject of Chapter 2). The lower the bit rate, the more sophisticated the underlying model will be, in general. Different kinds of coders are often classified according to the model(s) they are based on. Examples are "waveform coders" which aim at conveying the waveform as accurately as possible, "perceptual coders" which exploit the

properties of hearing by not conveying the precise waveform but only perceptually relevant parts of it, and "vocoders" which only transfer fully parameterized (speech production) information about the speech signal with the consequence that the original waveform cannot be reconstructed at all. In addition, all kinds of combinations of these classes of coders exist. Within these classes different kinds of compression methods are used, such as differential coding or coding on the basis of "analysis-by-synthesis" (see Chapter 3).

1.2.3 Bit rate

A primary parameter of a speech coder is the bit rate. It is often prescribed by the intended application. Sometimes this is a fixed bit rate but sometimes it can be a variable bit rate as well (see also Section 1.2.9 on scalability). In the case of a variable bit rate it may be relevant to know the minimum and maximum bit rates and also the rate of change of the bit rate. The format of the bit stream, i.e. the kind of information and its ordering within the bit stream, is also an issue. It can vary from one application to another and often formats are prescribed by standards.

1.2.4 Speech quality

Speech quality of a decoded signal is difficult to define since subjective issues like naturalness, noise and intelligibility are involved [9]. One objective measure is the signal-to-noise ratio (SNR). It is defined by the ratio of the signal energy and the noise energy, both measured over the same time interval. The "noise" is usually determined by the difference between the reproduced and original speech waveforms. The SNR only correlates well with subjective quality if it concerns relatively low-level noise and distortions. A somewhat better indication is the segmental SNR, which points out the average of the SNRs of a well-defined selection of time segments. An example of this is the particular selection of non-pause segments. It is clear that a bad SNR of hardly audible pauses in the speech signal cannot be a contribution to a relevant quality measure.

In non-waveform coders an SNR measure is completely useless. Also in waveform coders which utilize perceptual properties, notably those coders which take advantage of masking of the coding error by the speech signal itself, the SNR cannot be a reliable quality measure. The most extreme example of this is a coder which only inverts the signal. The SNR

MOS	Quality Indication	
5	Excellent	Commentary
4	Good	Toll
3	Fair	Communication
2	Poor	
1	Bad	

Table 1.1: MOS values and their associated quality indications.

of this coder equals -6 dB, while perceptually there is no difference at all between the two representations. More sophisticated objective measures than SNR, such as spectral distance measures, have been investigated and some of them even include masking models conform human hearing [10].

The paper on speech quality of Dimolitsas [11] gives a nice introduction to subjective assessment techniques that can be used to quantify the quality of speech. One method, which has already been in use for decades, is the Mean Opinion Score (MOS) test. In this method, a large number of listeners are asked to assess the quality of randomly sequenced utterances recorded from several speakers, using a 1-5 scale in terms of bad, poor, fair, good and excellent, respectively (see Table 1.1). After statistical processing of the results, an MOS number is obtained. Some of the speech material used may deliberately be contaminated by background noise or transmission errors, so that these issues are also included in the resulting MOS. In order to normalize these results so that they can be compared with other measurements, speech corrupted by what is called the Modulated Noise Reference Unit (MNRU) can be included in the tests [12]. In an MNRU noise is modulated by the speech signal and several test items are generated to cover a range of SNRs on the basis of this noise, so that the MOS values are paired to reproducible SNR values.

Some other widely used quality indicators which specify quality in a less precise way are "toll" quality which refers loosely to "normal" network telephone quality and which could be qualified as a speech quality with an MOS-value around 4. The indicator "communication" quality is typically found in the area of mobile telephony and digital answering machines and could be associated with a MOS-value around 3.5. On the other hand, "commentary" quality indicates a significantly better than

toll quality with an MOS-value around 4.5.

In addition to the so-called "absolute category rating" (ACR) 1-5 MOS-scale there is also a method which is based on "degradation category rating" (DCR) and the associated DMOS-scale ranges from 1 (very annoying) to 5 (not perceived). There exists also a method based on a "comparison category rating" (CCR) and its associated CMOS values range from 3 (much better) to -3 (much worse). A good overview of the several MOS related methods is also found in [13]. Several other application-dependent subjective measures exist, for instance, the "diagnostic acceptability measure" (DAM) [14] and the "diagnostic rhyme test" (DRT) [15]. While the former has a more elaborate scale than the MOS (16 scales of 0 to 100 points each), the latter aims at measuring the intelligibility alone using phonetically "minimal pairs" such as back/pack, which differ in only one phoneme.

1.2.5 Transparency

The transparency of a speech coder for other signals than speech signals can also play a role in the determination of the coder quality. In a coder which is based on a speech-specific model, signals that do not fit the speech model will be more distorted than signals that fit the model well. Examples of signals that certainly do not fit a speech-specific model are street noise, party noise and several kinds of music.

Also in the case of tandeming of codecs signals will occur which do not fit the speech-specific model. In tandeming, multiple codecs are cascaded, and the first codec will pass a signal to the second codec which is distorted and/or contains quantization noise. Even if the deterioration of the speech signal by the first codec is perceptually not relevant, the second codec may not be able to cope with the distortions so that the total chain will provide a lower speech quality. Tandeming may occur in mobile telephony, for instance, in the case of a mobile to mobile call, where each subscriber uses a codec in the radio link to his/her nearest base station.

1.2.6 Robustness

The robustness of a speech coder with respect to different speakers is also a quality issue. Some speakers fit the underlying model of a particular speech coder better than others. A speech coder can also show some sensitivity for different languages. The staccato sequence of altering

vowels and consonants in Japanese, for instance, generally results into better MOS scores for a coder under test, than other languages, such as some East-European languages, in which subtle differences in complex structured consonant clusters occur. The most important robustness measure, however, is often the sensitivity to transmission errors, at least in a communication environment. In storage applications transmission errors do generally not occur.

Transmission errors cause erroneous decoding. The decoder itself must be designed such that errors are perceptually minimized. Usual approaches serving this purpose include minimization of error propagation in the decoding process and minimization of the perceptual impact of single, isolated, bit errors.

1.2.7 Delay

The delay from the input of the encoder to the output of the decoder is an issue in full-duplex speech communications, as in telephony, because it may cause disturbing echoes. In a full-duplex telephone connection, speech received on a handset can easily enter the microphone on the same handset so that it is reflected to the transmitting party again. Sometimes, it is even necessary to employ expensive echo-cancellers, in which case minimization of the delay will help to reduce their costs. The delay requirements imposed on a speech coding system, often specified in terms of intrinsic "algorithmic delay" and hardware dependent "implementation delay", depend on the specific application and they vary from five to some tens of milliseconds. In a one-way GSM connection, for instance, the delay is about 90 ms including an algorithmic delay of 20 ms, 40 ms for error-protection measures (interleaving and channel coding), and the remainig delay is due to time division multiplexing and implementation delay. In a one-way connection between two locations on earth via a geostationary satellite - see Section 1.3.2 (under "maritime telephone") - the delay of the radio path alone amounts to approximately a quarter of a second. Such delays cause trouble in a conversation because the normal human reaction time in conversation is much faster. In half-duplex communications on the other hand, in which a communication channel is used in either direction at a time, the delay is not so much of an issue.

1.2.8 Complexity

The complexity of a speech coding system influences the battery lifetime, for instance, in a mobile telephone. Or, it determines whether a coder can run in real-time on a PC, or not. Usually, the complexity of a speech coder is of the order of the computing power of a modern single-chip digital signal processor (DSP) [16]. Complexity is often quantified in terms of MIPS, the number of millions of instructions per second. Another quantity is MOPS, million operations per second. A weighted MOPS measure, wMOPS is used by the ETSI [17], to indicate the complexity of its speech coding standards.

Systems with a high complexity will require more DSPs and systems with a low complexity can be realized using only a part of the computing capacity of one DSP. In general, a lower bit rate or a higher speech quality will require a higher complexity speech codec. Sometimes, the distribution of the complexity over the encoder and the decoder plays a role, such as in voice response systems in toys, amongst others. The encoder is implemented only once in such applications, while the decoder is made in manyfold. In this case it is important to keep the decoder as simple as possible, while this is not a prerequisite for the encoder.

1.2.9 Scalability

Scalability is a property which enables flexible use of a speech coder. It may concern different abilities. In one from of scalability the bit is controlled by the speech signal itself. Another form of scalability, which is sometimes referred to as bit-stream scalability, or as embedded coding or hierarchical coding, constitutes a layered bit stream with a core layer and one or more enhancement layers. This allows the omission of enhancement layers during transmission with the advantage of lowering the bit rate albeit at the cost of quality. This may be useful in network bottleneck situations where there is only the choice between lowering the bit rate or no transmission at all, or in cases that the same bit stream can be used for different receivers with different bit rates. This form of scalability is a central theme of the (ISO/IEC) MPEG-4 standard [18]. Another form of scalability is where the network gives feedback to the speech coder to adapt its bit rate. This is for instance applied in the GSM adaptive multi-rate (AMR) standard [19]. In this case the bit rate is adapted to the condition of the radio channel.

1.3 Applications of speech coding

The following survey of applications of speech coding is ordered according to the next subsections on (wired) telephony, mobile telephony, and storage applications. Speech coders are usually also available *on PC's*, however, sometimes as part of the operating system software, ready to be used in speech communications as well as in storage [45]. Whereas these applications all concern public access another subsection is added reviewing typical non-public applications.

1.3.1 Telephony

Even though experimental systems for *digital telephony* were studied quite soon after the invention of PCM by Alec H. Reeves in 1938, such as the 12-channel PCM system with toll quality reported in 1948 in [20], it lasted until the sixties of that century before digitalization of telephone exchanges took off on a large scale [21]. All incoming speech in such a telephone exchange can be digitalized and all outgoing speech can be transmitted in analog form again. Such a *digital exchange* has the advantage that it, once and for all, solves the problems of the cumbersome electro-mechanical switches of analog exchanges. Such a strategy has the disadvantage, however, that the speech quality suffers from tandem connections of codecs of the exchanges through which the speech signal passes on its way from one subscriber to the other. *Digital transmission* between the exchanges avoids this and in addition it creates better possibilities for signalling and multiplexing of many voice channels on a single physical medium (time division multiplex - TDM - which is much more efficient than the traditional analog frequency division multiplex - FDM - requiring a modulation and demodulation stage for each voice channel). Moreover, digital transmission is capable of delivering a speech quality which is independent of the distance between the end points of the connection. This results from the usage of repeaters which regenerate the signal at regular distances before it is deteriorated too much by attenuation and noise. Thus, it can happen that a digital telephone connection with (ITU-T) standard PCM at 64 kbit/s between for instance Europe and the United States sounds better than an analog connection on a long local line. An elaborate treatise of digital telephony is given in the book of Bellamy [22].

On busy and expensive connections such as *transatlantic cables* and *satellite links*, lower bit rates than 64 kbit/s may be desirable or some-

times even required. There are systems in operation such as TASI (time assignment speech interpolation) [23], also referred to as digital speech interpolation (DSI) [24], for *digital circuit multiplication* equipment (DCME) to increase the efficiency of such connections. A TASI system is a dynamic multiplexer which divides a large number of calls over a smaller number of connections by utilizing the speech pauses in the conversations. Traditionally this resulted in a gain of a factor of 2.5. By utilizing (ITU-standard) speech compression to a bit rate of 32 kbit/s, the gain is almost doubled [25]. Nowadays, many international calls over cables and satellites are compressed.

By the use of digital technology it is straight-forward to multiplex different kinds of signals, such as speech, image and control or other data (*multimedia*), and to transmit them over a single communication medium. At the receiving side they can easily be separated again. One of the applications which makes use of such possibilities is *video conferencing*. Sometimes, broadband connections are available for this purpose, but not always [26,27]. It is also possible to use an ISDN (integrated services digital network) connection which provides a transmission capacity of 128 kbit/s [28]. It is often preferred to allocate a bit rate as low as possible to the speech signal, in this case, in order to maximize the bit rate for the video signal. For this purpose, ITU-T standard speech coders are available at various bit rates down to 5.3 kbit/s [29].

The latest development is telephony over packet-switched networks on the basis of the Internet protocol IP [33,34]. Sometimes, *Voice over IP (VoIP)* makes use of the existing infrastructure of the Internet. The usual charges for telephony do not apply here and in principle one can call all over the world against local rates, or in any case against the costs of using the Internet. Also for the network operator IP telephony is an attractive option because packet-switched networks can be utilized better than the conventional circuit-switched networks. However, the Internet has some limitations which have a negative impact on real-time full-duplex speech communication. These limitations are unpredictable delays which cause so-called jitter, the loss of packets, and often still insufficient available bit rate on the local subscriber lines of the net. These transmission problems can be alleviated somewhat by using low bit rate speech coders, but in order to obtain good speech communication a guaranteed minimum "quality of service" (QoS) is necessary. Such a functionality can give a certain priority to speech packets in the network, so that delivery to the recipient within a certain time can

be guaranteed. This will undoubtedly have its impact on the tariffs. Another issue is the overhead of IP packet headers which may even exceed the VoIP-payload by a factor of two. Increasing the payload per packet (with less packets) increases the delay, however. Another option is header compression. It is not yet fully clear where VoIP is leading to, but it is clear by now that it is here to stay.

1.3.2 Mobile telephony

In the nowadays very popular *mobile telephony*, digitalization has caused a real revolution. Besides the good adjacent channel interference[1] properties on the radio path, the good co-channel interference[2] properties of a digital solution allow frequencies to be reused at relatively smaller distances in the cellular infrastructure of the system as compared to analog solutions. Therefore, digital solutions provide for enhanced spectral efficiency so that more users per square kilometer can be served as compared to the out-of-date analog solutions [35]. In addition, the digital bit stream can be better protected against the adverse environment of a radio channel, causing e.g. multi-path fading, by making use of advanced digital modulation methods and channel coding. It also enables the application of *cryptography* making eavesdropping very difficult. Moreover, digital solutions pave the way to the application of other (multimedia) services such as mobile Internet access.

In European and many other countries of the world the (ETSI) GSM system is used [36,37], see Table 1.2[3], while in the USA and Japan other systems are used. The first standard speech coder for digital mobile telephony, which is still used today, was the so-called full-rate speech coder of the GSM system, described in the core-chapter of this thesis (Chapter 4). The bit rate for speech coding in a full-rate GSM-channel is 13 kbit/s, while the gross bit rate of the channel is 22.8 kbit/s. The rest is used for error protection. Although toll quality at such bit rates is possible today, this coder was designed in the early days of digital mobile telephony in the 1980s and the average performance is limited to communication quality. The main cause of inconvenience is not this

[1] The radio system may include an FDM architecture, so that adjacent to the used FDM channel another channel may cause interference.

[2] The channel in an FDM radio system at the same frequency but at another, not too distant, geographic location may also cause interference.

[3] Table 1.2 gives a total number of subscribers of 872 491 220 by the end of June 2003. According to the website mentioned in reference [37], this number has grown to 1.4 billion by August 2005.

Region or Continent	Number of Countries	Subscribers End June 03	Penetration in %
Africa	42	42 005 550	4.94
USA/Canada	2	21 725 730	6.89
Other Americas	28	12 302 230	2.42
Asia Pacific	35	338 414 320	10.17
Europe	59	424 771 210	48.52
Middle East	13	19 900 180	10.8

Table 1.2: Some GSM statistics from 2003(after [37], see also footnote 3).

quality limitation, however, but the occasional dropouts of the radio channel. In the design of the speech coder and the accompanying error protecting channel coder special attention has been paid to optimize their performance for the adverse radio channel. If the performance of the radio channel drops below a certain threshold, however, the speech signal will be faded out.

A different kind of mobile telephone service is the *maritime telephone* meant for small vessels, known as Inmarsat-M [38,39]. This service makes use of geostationary satellites. The height of the orbit of such satellites is about 36000 km above the earth. Because the dimensions of the antenna are limited to the order of magnitude of one meter for the relatively small vessels, for instance for a disk antenna, and because the available power in the satellite as well as in the vessel is limited, the communication capacity is confined to a few kilobits per second [40]. Because of the satellites this service is worldwide available and it only exists thanks to advanced low-bit-rate speech coding. Better link energy budgets are possible with satellites in lower orbits, such as in the Iridium system [41].

The popular *cordless telephone* handset for the "plain old telephone service" (POTS) has analog as well as digital variants. One of the digital versions is made according to the "digital European cordless telephone" (DECT) ETSI-standard and it uses speech coding at 32 kbit/s [42].

1.3.3 Storage

Speech coders enable efficient solutions to high quality *storage* applications such as *voice mail*. In a voice mail or voice message system a recorded message is stored on a server somewhere in the network and

it can be listened to at a later time, exclusively by the intended recipient. Another application is *voice response* by which in reaction to an event, such as pressing a key on a telephone pad, certain pre-recorded announcements can be rendered [30,31,32].

Speech coding is also applied in personal speech *recording* equipment, such as digital telephone *answering machines* [43] and digital *pocket memo's*. These devices use non-volatile flash memory in the storage function. It is a good economic trade-off to apply a speech coder in order to minimize the memory size.

Still other applications of speech coding are in the field of pre-recorded ("canned") speech. Speaking *toys*, *public address* systems with a number of pre-recorded messages - such as in *car navigation* systems - and *spoken books* stored on tape or CD - which are quite popular in the USA - belong to this class. Already in 1978, Texas Instruments introduced a battery-operated toy called "Speak & Spell" [44].

1.3.4 Non-public applications

Besides the public applications there also exist *non-public speech communication applications*. They are mainly found in military applications, space communications and in wired as well as mobile private networks.

In *military applications* it is especially important that digital speech can be protected against eavesdroppers by the application of cryptography [46]. *Secure speech* is also used by other government institutions (*intelligence, judiciary*) and even by private organizations (*business*). In the military applications sometimes own networks are employed such as the NATO Deltamux system from the 1980s [47]. This system operates at bit rates of 32 and 16 kbit/s. The public telephony infrastructure is used as well for this purpose. In this case less than 4 kHz bandwidth is available for the transmission of the digital signal making speech coding at low bit rates of 2.4 or 4.8 kbit/s a prerequisite. The youngest secure voice system of the USA Department of Defense (DoD), which can also be used in wireless applications, was announced in 1996 and it uses a bit rate as low as 1.6 kbit/s [48]. In wireless secure speech communication systems it can occur that the system has to cope with *degraded channels* with a very bad signal to noise ratio. By applying speech coding at a very low bit rate of a few hundred bits per second it becomes possible to use heavy channel coding with much redundancy so that useful communication can yet be obtained [46].

In *space* long distances play a role which translate into combating

noise. As a result of the possibility to take binary decisions in the regeneration process in the receiver in combination with the use of advanced channel coding, digital transmission is preferred to analog transmission. An example is the space-shuttle coder described in [49].

In private networks, switching takes place in so-called *Private (Automatic) Branch Exchanges* (PABX/PBX). The connections between the exchanges and from the private network to the PSTN can carry more voice channels using compression techniques, providing better solutions from an economical point of view. As early as in 1979 [50], a speech digitizer at 2400 bits/s, called TSP-100, was recommended to be used to obtain four conversations simultaneously on a single line, or to mix speech with other digital data. In at least one example the PABX is combined with cordless technology [51].

In *Private Mobile Radio* (PMR) the speech traffic is multiplexed over a limited number of especially allocated radio channels. The equipment is owned by the users themselves. Typical owners are e.g. *police, fire brigades, ambulance* services and *taxi* companies. In modern digital PMR systems the spectral efficiency (the number of simultaneous users per frequency band) is higher as compared to analog systems [52], and secure speech can be provided by encryption. The ETSI Trans-European Trunked Radio (TETRA) system uses speech coding at a bit rate of 4.56 kbit/s [53].

1.4 State of the art in speech coding

The state of the art in speech coding is depicted in Figure 1.2. It shows the speech quality of several - narrow-band - standard coders and their bit rates. The names of the various coding methods mentioned in this section are included for informative reasons only. They will not be explained here. At 64 kbit/s there is the ITU-T Recommendation G.711 entitled "Pulse code modulation (PCM) of voice frequencies", originating from 1972. Extensive background information about it is found in the book of Jayant and Noll [3]. The transparency of this simple, universal, coder for non-speech signals is very good and the coding delay is zero.

At 32 kbit/s there is the ITU-T coder according to Recommendation G.726: "40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (ADPCM)". This recommendation combines and replaces - since 1990 - Blue Book Recommendations G.721 (32 kbit/s, 1984) and G.723 (exten-

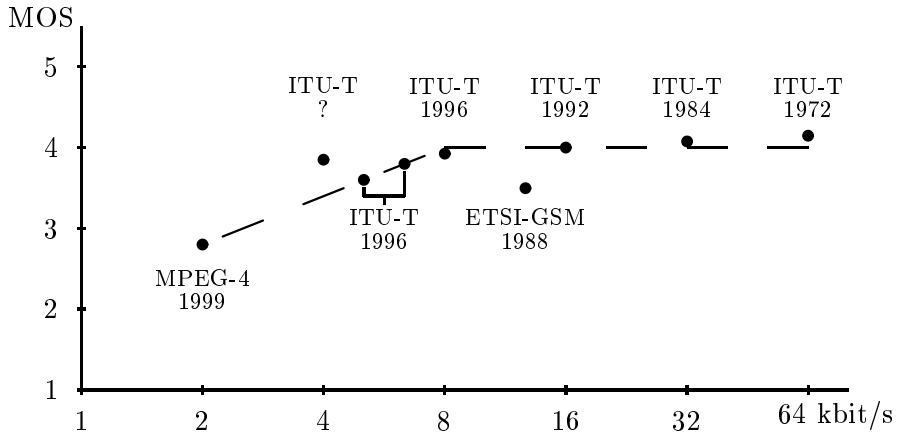

Figure 1.2: State of the art in speech coding.

sion to 24 and 40 kbit/s). Detailed technical elaborations on this coder are found in another chapter of the same book [54]. This coder has also no delay and it is somewhat more complex and a little less transparent than PCM, but on the other hand it provides for bit rate scalability which is especially meant for overload situations in DCME applications.

The speech coder according to ITU-T Recommendation G.728 from 1992: "Coding of speech at 16 kbit/s using low-delay code excited linear prediction" has already lost some performance concerning several attributes. Especially designed for low delay, the one-way end-to-end delay of the codec - including an algorithmic delay of 0.625 ms - amounts to a modest 2 ms. The transparency is still acceptable, but the sophisticated algorithm is very complex [55]. This means that the algorithm asks for specially designed ASIC (application specific integrated circuit) realizations. The sensitivity for transmission errors is also an issue, but under normal fixed-network conditions this is not yet a problem.

The state of the art at 8 kbit/s is given by the ITU-T Recommendation G.729 from 1996: "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code excited linear-prediction (CS-ACELP)". This coder cannot reach the level of transparency of the coders at the higher bit rates and non-speech signals are distorted, as a result. The algorithmic delay amounts to 15 ms and a 1% bit error rate is allowed to cause a 0.5 MOS performance decrease from error-free G.726 at 32 kbit/s [56]. The considerable complexity of the algorithm still allows implementa-

tion onto a single general purpose DSP (1996-technology).

Also from 1996, there is the G.723.1 ITU-T coder with the title "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s" [57]. This coder was especially designed for videophone applications on the PSTN (as part of the overall H.324 family of standards). This CELP (code excited linear prediction) coder provides some scalability by the dual bit rate. The low bit rates are made possible by allowing an algorithmic delay as high as 37.5 ms. The transparency of the coder corresponds at most to that of the G.729 and the lower bit rate of 5.3 kbit/s cannot provide toll quality anymore.

While extending the trend of the 1990s - as shown by the right-hand, horizontal, part of the dashed line in Figure 1.2 - to lower bit rates, the ITU-T has called for 4-kbit/s proposals providing toll quality, in 1994 already [58]. Up till now (early 2005), this has not yet led to a complete solution which fulfils all requirements, but it is not unlikely that such a solution will be created in the near future [59].

At the very low bit rate end of the curve at 2 kbit/s, the state of the art is very well represented by the (ISO/IEC-)MPEG-4 [18], fully parametric, coder HVXC (harmonic vector excitation coder), which is based on a sinusoidal coding model [60]. Due to its reliance on a very speech-specific model, its performance is only acceptable for (clean) speech signals. The delay of this vocoder is several tens of milliseconds. Because of its parametric nature, this vocoder offers the possibility for time- and frequency-scale modifications of speech.

The MPEG-4 standard, set in 1999, has scalability as a central theme [18]. In the speech coding part a scalable CELP coder with MPE (multi-pulse excitation) is standardized for bit rates ranging from 4 to 12 kbit/s. In addition, wide-band coding is covered on the basis of a CELP with MPE as well as RPE (regular pulse excitation), for the bit rate range of 11-24 kbit/s. These narrow-band and wide-band codecs are not shown in Figure 1.2. CELP, MPE and RPE are explained in Chapter 3.

Figure 1.2 also shows the 13 kbit/s full-rate GSM coder - ETSI Recommendation 06.10: "GSM full rate speech transcoding", from 1988 - which is described in Chapter 4. It stands out in that the bit rate of this coder is quite low according to the state of the art of the time (mid 1980s) while communications quality is still provided. This is due to the application of RPE and LTP (long-term prediction). The strength of this coder lies in the combination of these attributes with an acceptable algorithmic delay of 20 ms, a good robustness with respect to

transmission errors, a reasonable transparency and, last but not least, a convenient low complexity. The coder had only two predecessors - the 1972 and the 1984 ITU-T standards - and it was the first speech coder ever standardized for public digital mobile telephony [61].

1.5 Organization

The contents of this thesis is organized in the following way.

Chapter 2 starts with a summary of the features of speech signals and their source, the human speech organ. Then, historical models of speech production which form the basis of different kinds of speech coders are discussed. Starting with a review of ancient mechanical models, we will arrive at the electrical source-filter model of the 1930s. Subsequently, the acoustic tube models as they arose in the 1950s and 1960s are discussed. Finally the 1970s are reviewed which brought the discrete-time filter model on the basis of linear prediction. In a unique way the logical sequencing of these models is exposed, and the links are discussed. Whereas the historical models are discussed in a narrative style, the acoustic-tube models and the linear prediction technique as applied to speech, are subject to more mathematical analysis in order to create a sound basis for the treatise of Chapter 4. This trend continues in Chapter 3, whenever instrumental in completing that basis.

In Chapter 3 the reader is taken by the hand on a guided tour through time during which successive speech coding methods pass in review. In an original way special attention is paid to the evolutionary aspect. Specifically, for each newly proposed method it is discussed what it added to the known techniques of the time. After presenting the relevant predecessors starting with PCM and the early vocoders of the 1930s, we will arrive at RELP (Residual-Excited Linear Predictive) coders, Analysis-by-Synthesis systems and Regular-Pulse Excitation in 1984. The latter forms the blue-print of the GSM full-rate coder.

In Chapter 4, which constitutes the core of this thesis, explicit forms of Multi-Pulse Excited (MPE) and Regular-Pulse Excited (RPE) analysis-by-synthesis coding systems are developed. Starting from current pulse-amplitude computation methods in 1984, which included solving sets of equations (typically of order 10-16) two hundred times a second, several explicit-form designs are considered by which solving sets of equations in real-time is avoided. Then, the design of a specific explicit-form RPE coder and an associated efficient architecture are described.

The explicit forms and the resulting architectural features have never been published in so much detail as presented here. Implementation of such a codec enabled real-time operation on a state-of-the-art single-chip digital signal processor of the time. The selection of this coder, at a bit rate of 13 kbit/s, as the Full-Rate GSM standard in 1988 and its performance are recapitulated.

Chapter 5 is a brief epilogue reviewing the relevant developments in speech coding technology after 1988. Many speech coding standards have been set, for mobile telephony as well as for other applications, since then. The chapter is concluded by an outlook.

References

1. H. Nyquist, Certain Topics in Telegraph Transmission Theory, *Proceedings of the IEEE*, Volume 90, No.2, February 2002, pp. 280–305. Reprinted from *Transactions of the A.I.E.E.*, February 1928, pp. 617–644.
2. B.M. Oliver, J.R. Pierce and C.E. Shannon, The Philosophy of PCM, *Proceedings of the IRE*, November 1948, pp. 1324–1331.
3. N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1984, Chapter 5: Pulse Code Modulation, pp. 221–251.
4. M.H. Sherif and D.K. Sparrell, Standards and Innovations in Telecommunications, *IEEE Communications Magazine*, July 1992, pp. 22–29.
5. T. Irmer, Shaping Future Telecommunications: The Challenge of Global Standardization, *IEEE Communications Magazine*, January 1994, pp. 20–28.
6. G. Robin, The European Perspective for Telecommunication Standards, *IEEE Communications Magazine*, January 1994, pp. 40–44.
7. S. Dimolitsas, Standardizing Speech-Coding Technology for Network Applications, *IEEE Communications Magazine*, November 1993, pp. 26–33.
8. Specification for an intermediate reference system, *ITU-T Recommendation* P.48.
9. N. Kitawaki and H. Nagabuchi, Quality assessment of speech coding and speech synthesis systems, *IEEE Communications Magazine*, October 1988, pp. 36–44.
10. J.G. Beerends, Audio quality determination based on perceptual measurement techniques, Chapter 1 in: M. Kahrs and K. Brandenburg (Eds.), *Applications of digital signal processing to audio and acoustics*, Kluwer Academic Publishers, 1998, pp. 1–38.
11. S. Dimolitsas, Subjective assessment methods for the measurement of digital speech coder quality, in: B. Atal, V. Cuperman and A. Gersho (Eds.), *Speech and Audio Coding for Wireless and Network Applications*, Kluwer Academic Publishers, 1993, pp. 43–53.
12. Modulated noise reference unit, *ITU-T Recommendation* P.810.
13. P. Kroon, Evaluation of Speech Coders, Chapter 13 in: W.B. Kleijn and K.K.Paliwal (Eds.), *Speech Coding and Synthesis*, Elsevier Science B.V., Amsterdam, 1995, pp. 467–494.

14. W.D. Voiers, Diagnostic acceptability measure for speech communication systems, *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Hartford, 1977.
15. W.D. Voiers, Diagnostic evaluation of speech intelligibility, in: M.E. Hawley (Ed.), *Speech Intelligibility and Speaker Recognition*, Dowden, Hutchinson & Ross, Stroudsberg, PA (John Wiley & Sons, Inc.), 1977.
16. J.L. Flanagan, Speech Technology and Computing: A Unique Partnership, *IEEE Communications Magazine*, May 1992, pp. 84–89.
17. Complexity and delay assessment, AMR Permanent Document (AMR-9), *ETSI SMG11 AMR#9, Versailles, France*, 23-25 March, 1998.
18. K. Brandenburg, O. Kunz and A. Sugiyama, MPEG-4 Natural Audio Coding, *Signal Processing: Image Communication*, Tutorial Issue on MPEG-4 Standard, Volume 15, Nos.4-5, January 2000, pp. 423–436.
19. S. Bruhn, P. Bloecher, K. Hellwig and J. Sjoeberg, Concepts and solutions for link adaptation and inband signaling for the GSM AMR speech coding standard, *IEEE 49th Vehicular Technology Conference*, 1999, Volume 3, pp. 2451–2455.
20. L.A. Meacham and E. Peterson, An Experimental Multichannel Pulse Code Modulation System of Toll Quality, *The Bell System Technical Journal*, January, 1948, pp. 1–43.
21. M.R. Aaron, Digital Communications - The Silent (R)evolution?, *IEEE communications Magazine*, January 1979, pp. 16–26.
22. J. Bellamy, *Digital Telephony*, John Wiley & Sons, New York, 1991.
23. E.F. O'Neill, TASI, *Bell Laboratories Record*, March 1959, pp. 82–87.
24. Special Issue on Bit Rate Reduction and Speech Interpolation, *IEEE Transactions on Communications*, April 1982.
25. M.H. Sherif, The Metamorphosis of the Public Switched Telecommunications Network, *IEEE Communications Magazine*, January 1994, pp. 14–16.
26. J.D. Gibson (Ed.), *Multimedia Communications, Directions and Innovations*, Academic Press, San Diego, 2001, Chapter 1: Mul-

timedia communications: Source Representations, Networks, and Applications, pp. 1–12.

27. L.F.M. de Moraes and S.B. Weinstein, The Internet Multicast from ITS: How it was Done and Implications for the Future, *IEEE Communications Magazine*, January 1995, pp. 6–8.
28. M.L. Liou, Visual Telephony as an ISDN Application, *IEEE Communications Magazine*, February 1990, pp. 30–38.
29. R.V. Cox and P. Kroon, Low Bit-Rate Speech Coders for Multimedia Communication, *IEEE Communications Magazine*, December 1996, pp. 34–41.
30. L.R. Rabiner, Applications of Voice Processing to Telecommunications, *Proceedings of the IEEE*, February 1994, pp. 199–228.
31. E.A. Munter, Digital Switch Digitalks, *IEEE Communications Magazine*, November 1982, pp. 15–23.
32. P. Mermelstein, Voice Message Systems, *IEEE Communications Magazine*, December 1983, pp. 8–10.
33. Special Issue on Internet Telephony, *IEEE Network*, May/June 1999.
34. O. Hersent, D. Gurle, J-P. Petit, *IP Telephony (Packet-Based Multimedia Communications Systems)*, Pearson Education Limited, Edinburgh, 2000.
35. J.E. Natvig, S. Hansen and J. de Brito, Speech Processing in the Pan-European Digital Mobile Radio System (GSM) - System Overview, *IEEE Global Telecommunications Conference (Globecom)* 1989, Dallas, paper 29B.1, pp. 1060–1064.
36. S.M. Redl, M.K. Weber, M.W. Oliphant, *An Introduction to GSM*, Artech House Inc., Norwood MA, 1995.
37. *3GSM World Focus* 04, Statistics, pp. 125–136, Informa Business Publishing Ltd., London.
    See also www.gsmworld.com/news/statistics
38. J.C. Hardwick and J.S. Lim, The Application of the IMBE Speech Coder to Mobile Communications, *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 1991, pp. 249–252.
39. S.F. Campos Neto, F.L. Corcoran, J. Phipps and S. Dimolitsas, Performance Assessment of 4.8 kbit/s AMBE Coding under Aeronautical Environmental Conditions, *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 1996, pp. 499–502.

40. J.H. Lodge, Mobile Satellite Communications Systems: Toward Global Personal Communications, *IEEE Communications Magazine*, November 1991, pp. 24–30.
41. B. Miller, Satellites free the mobile phone, *IEEE Spectrum*, March 1998, pp. 26–35.
42. W.H.W. Tuttlebee (Ed.), *Cordless Telecommunications in Europe*, Springer-Verlag, Berlin, 1990.
43. W. Armbruester, S. Dobler and P. Meyer, Hands-free telephony, speech recognition and speech coding techniques implemented in the SPS51, *Philips Telecommunication Review*, March 1991, pp. 19–27.
44. R. Wiggins, An integrated circuit for speech synthesis, *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Denver, CO, 1980, pp. 398–401.
45. H. D'Hooge, The Communicating PC, *IEEE Communications Magazine*, April 1996, pp. 36–42.
46. C.J. Weinstein, Opportunities for Advanced Speech Processing in Military Computer-Based Systems, *Proceedings of the IEEE*, November 1991, pp. 1626–1641.
47. J.W. Glasbergen, Second generation deltamux, *Philips Telecommunication Review*, September 1985, pp. 193–201
48. A. McCree and J.C. De Martin, A 1.6 kb/s MELP Coder for Wireless Communications, *IEEE Workshop on Speech Coding for Telecommunications Proceedings*, Poco Manor, Pennsylvania, USA, 1997, pp. 23–24.
49. R.L. Auger, M.W. Glancy, M.M. Goutmann and A.L. Kirsch, The Space Shuttle Terminal Delta Modulation System, *IEEE Transactions on Communications*, November 1978, pp. 1660–1670.
50. S. Maitra and C.R. Davis, A Speech Digitizer at 2400 Bits/s, *IEEE Transactions on Acoustics, Speech and Signal Processing*, Volume ASSP-27, No.6, December 1979, pp. 729–733.
51. C. Buckingham, G. Klein Wolterink and D. Akerberg, A Business Cordless PABX Telephone System on 800 MHz Based on the DECT Technology, *IEEE Communications Magazine*, January 1991, pp. 105–110.
52. W. Holubowicz, 1990's - The Decade of Pan-European Digital Standards in Wireless Communications, *International Conference on Personal Wireless Communications*, 1996, pp. 91–95.

53. P. Whitehead, The Other Communications Revolution, *IEE Review*, July 1996.
54. N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1984, Section 6.5.3: *DPCM-APB-AQB for Toll Quality Speech at 32 kb/s*, pp. 307–311.
55. J.-H. Chen, R.V. Cox, Y.-C. Lin, N. Jayant and M.J. Melchner, A Low-Delay CELP Coder for the CCITT 16 kb/s Speech Coding Standard, *IEEE Journal on Selected Areas in Communications*, June 1992, pp. 830–850.
56. R. Salami, C. Laflamme, J.-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon and Y. Shoham, Design and Description of CS-ACELP: A Toll Quality 8 kb/s Speech Coder, *IEEE Transactions on Speech and Audio Processing*, March 1998, pp. 116–130.
57. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, *ITU-T Recommendation* G.723.1
58. S. Dimolitsas, C. Ravishankar and G. Schroeder, Current Objectives in 4-kb/s Wireline-Quality Speech Coding Standardization, *IEEE Signal Processing Letters*, November 1994, pp. 157–159.
59. J. Thyssen, Y. Gao, A. Benyassine, E. Shlomot, C. Murgia, H. Su, K. Mano, Y. Hiwisaki, H. Ehara, K. Yasunaga, C. Lamblin, B. Kovesi, J. Stegmann, and H.-G. Kang, A Candidate for the ITU-T 4 kbit/s Speech Coding Standard, *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Salt Lake City, Utah, 2001.
60. M. Nishiguchi, A. Inoue, Y. Maeda, J. Matsumoto, Parametric Speech Coding - HVXC at 2.0-4.0 kbps, *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, 1999, pp. 84–86.
61. J. E. Natvig, Evaluation of Six Medium Bit-Rate Coders for the Pan-European Digital Mobile Radio System, *IEEE Journal on Selected Areas in Communications*, February 1988, pp. 324–331.

Chapter 2

Historical speech production models

This chapter starts with an introduction to the features of speech signals and to their source, the human speech organ. Then, historical models of speech production which form the basis of different kinds of speech coders are discussed. Starting with a revisit of ancient mechanical models, we will arrive at the electrical source-filter model of the 1930's. Subsequently, the acoustic tube models as they arose in the 1950s and 1960s are discussed. Finally the 1970s are reviewed which brought the discrete-time filter model on the basis of linear prediction. The originality of this chapter lies in the exposure of the logical sequencing of these developments and in the discussion of how they link. Whereas the historical models are discussed in a narrative style, the acoustic tube models and the technique of linear prediction as applied to speech (from Section 2.3.5 onwards), are subject to more mathematical analysis in order to create a sound basis for the treatise of Chapter 4. This trend continues in Chapter 3, whenever instrumental in completing that basis.

2.1 Speech signals and models

In this section we review the most important features of the human speech organ, speech signals, and their spectral representation, as well as an introduction to models of the speech production process.

2.1.1 The human speech organ

The characteristic properties of a speech signal can be explained with the aid of some insight in the physiology of the human speech organ. Figure 2.1 shows a sketch of the most important elements of the speech organ. Air pressure in the trachea 1 (windpipe), built up by the lungs, is the energy source of the system. It can produce different sounds which can be classified as either voiced or unvoiced, or the combination of both.

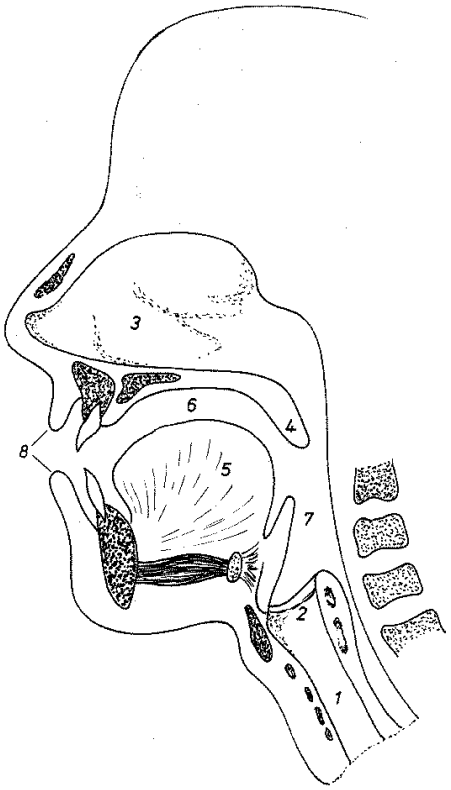

Figure 2.1: Cross-section of the human speech organs. 1 Windpipe or trachea, 2 vocal cords, 3 nasal cavity, 4 soft palate or velum, 5 tongue, 6 mouth cavity, 7 throat cavity or pharynx, 8 lips.

During a voiced sound the elastic vocal cords 2 vibrate under influence of the air pressure in the windpipe and the muscular tension on

them. These dynamic quantities as well as the more static parameters such as shape, mass and compliance of the vocal cords, will determine their vibration frequency. The vibration frequency is called *pitch* (in the perception literature pitch is a perceptual term, while in the speech coding literature pitch is a physical quantity). For children's voices the pitch can get as high as 500 Hz. In singing, a sopranos voice can even reach 1.5 kHz. In some adult male voices, on the other hand, the pitch can drop down to 50 Hz, especially at the end of a sentence. The periodic pulses of airflow generated by the vocal cords are rich in harmonic components and they excite the vocal tract acoustically over a relatively wide frequency range. The vocal tract ranges from the vocal cords up to the lips 8 and the nostrils and it further comprises the pharynx 7 (throat cavity), the mouth cavity 6 and the nasal cavity 3. Resonances in these cavities determine the transfer function of the vocal tract and as such, the timbre of the sound radiated from the mouth and the nostrils. These resonances are called *formants*. By varying the physical shape of the vocal tract the transfer function is varied. The elements which play an important role in this process are the tongue 5, the mouth opening between the teeth, and the position of the lips 8. The opening to the nasal cavity can be closed by the velum (soft palate) 4. This is the case during the production of non-nasal sounds [1]. During the production of nasal sounds - such as the consonants m, n and ng, according to [2] - the velum is open and the passage through the mouth cavity and lips is closed. This means that for most speech sounds only one tube is effectively involved, either from the vocal cords to the lips or from the vocal cords to the nostrils. The average length of the vocal tract of adult males amounts to approximately 17 cm.

During the production of unvoiced sounds the vocal cords do not vibrate. In this case the excitation is caused by an airflow which is brought into turbulence by a constriction somewhere in the vocal tract, such as for instance at the front teeth during the production of "s". As a result, a noise-like sound is produced and the transfer function of that part of the vocal tract which is located between the constriction and the lips determines the timbre of the sound, in addition to the spectral colour of the excitation signal itself.

There are also sounds which are generated by a combination of voiced and unvoiced excitation. An example of such a case is the consonant "z" as in zero. There exist several other classifications of speech sounds such as voiced plosives (or stops) "b", "d" and "g" and their respective

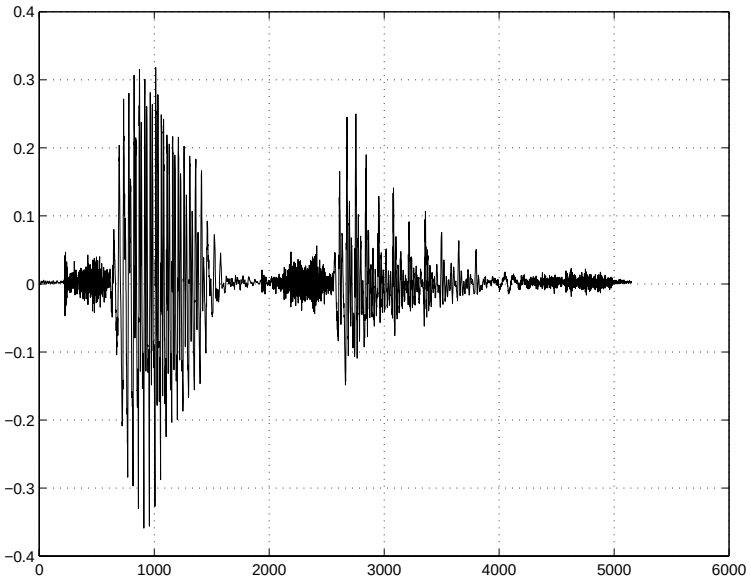

Figure 2.2: Waveform of the word "coaches". The horizontal axis shows sample numbers. The sampling frequency is 8 kHz.

unvoiced counterparts "p", "t" and "k".

Many models of speech production which one encounters in the literature consist of imitations of the speech production mechanism. Many, but not all, as we will see. Before some famous models will be discussed, however, we will investigate some real speech.

In Figure 2.2 the waveform of the word "coaches" is depicted which is cut from a normally uttered sentence spoken by a male speaker. The waveform has been sampled at 8 kHz and the time axis is represented in number of samples. The duration of this word is approximately 0.6 seconds. The vertical axis represents amplitude on an arbitrary linear scale. The voiced vowels with the quasi-periodic impulsive excitations and the much weaker, but over time more evenly distributed, noisy unvoiced consonants are clearly distinguishable. In these 0.6 seconds two different vowels are thus produced embedded between three different unvoiced consonants. The large difference in level of almost 30 dB between the maximum level of the first vowel and the level of the final "s"-sound is striking. The high dynamics of the pitch over this short time span is also astonishing. The shortest pitch period is located at the beginning

of the first vowel and amounts to 38 sampling periods (210 Hz). The longest pitch period is located at the end of the last vowel and amounts to 152 sampling periods (52.5 Hz). This is a range of two octaves which corresponds to the normal range of a singing-voice!

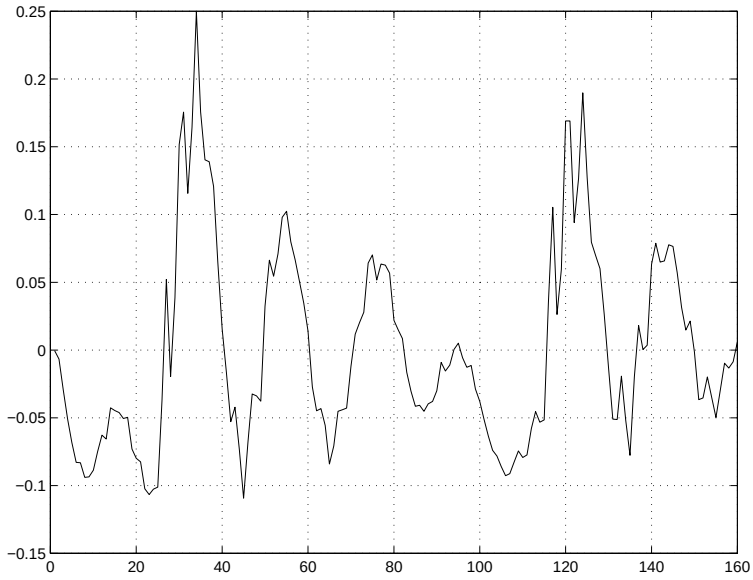

Figure 2.3: Segment of 160 samples (20 ms) from the last vowel in "coaches".

2.1.2 Spectral features of speech

Figure 2.3 shows a magnified segment of the second vowel "e". The time duration of this segment amounts to 160 sampling periods, or 20 ms. From the amplitudes it can be seen from which part of the waveform of Figure 2.2 the segment has been taken. The excitation instants and the decaying responses of the vocal tract are clearly visible in this segment. The pitch period amounts to 89 sampling periods which corresponds to approximately 90 Hz.

Figure 2.4 shows the points 0-400 of the 800-point FFT magnitude spectrum of this time function. For that purpose, 640 zero-valued samples are appended to the 160 samples of the segment. In this way a 401-point scale (0-400) is obtained which corresponds to a frequency range of 0 - 4 kHz. Because of the finite-duration window applied to the

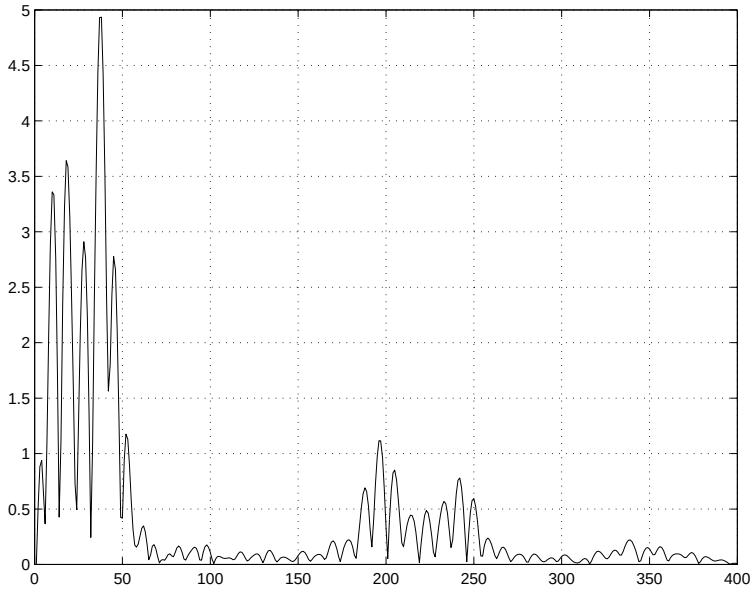

Figure 2.4: FFT magnitude spectrum of "e". The frequency points 0 - 400 of the horizontal axis correspond to a frequency range of 0 - 4 kHz.

time function a single sinusoidal component is represented in the spectrum by the Fourier transform of the window which has the shape of a lobe. In the spectrum we clearly see a fundamental frequency and the harmonically located lobes. In addition, several formants are manifest as groups of harmonics with a relatively high amplitude.

In the same way, an unvoiced segment and its FFT magnitude spectrum are determined from the final consonant "s". Figure 2.5 shows the noisy time-segment and Figure 2.6 shows its magnitude spectrum. In this spectrum, two formants manifest themselves, a sharp formant in the vicinity of 2500 Hz and a somewhat stronger but also wider formant around 3500 Hz. The zero-crossings in the time function appear to correspond roughly to the zero-crossings of a 3500 Hz sinusoid.

2.1.3 Models of speech production

The signals and spectra in the illustrations are representations of normal speech signals. On the one hand, the signal properties vary dynamically. On the other hand, the whole system is a muscle-controlled machine

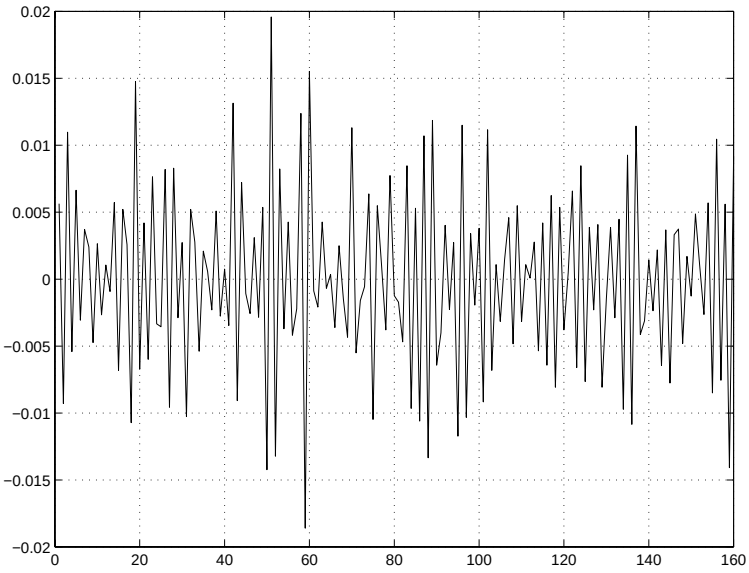

Figure 2.5: Segment out of the final unvoiced sound "s".

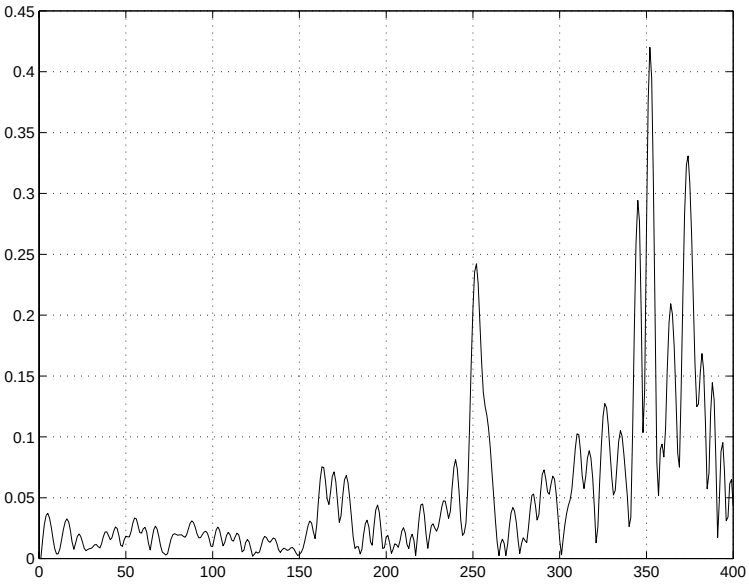

Figure 2.6: FFT magnitude spectrum of the "s" from Figure 2.5.

which consequently has its limitations in operational speed. As a compromise, we may consider the signal to be stationary over a short time interval, of typically 10-20 ms, during which parameters, such as pitch, formant locations and sound intensity, are representative for the speech signal at that time.

A logically resulting model of speech production from the preceding characterisation of the speech organ and the speech signal is the so-called *source-filter model*. In this model the excitation function is represented by a periodic impulse source (with a certain pitch) for voiced sounds or a noise source for unvoiced sounds. Either of these sources excites a filter with a transfer function such that the produced signal obtains the required spectral features. The model is completed by a parameter for the sound intensity.

The modelling of the speech organ and the generation of artificial speech have captivated scientists in the 18^{th} century, already. In those days there was, in several parts of Europe, an atmosphere of scientific curiosity and interest in experimenting with mechanical systems and in building automatons, often activated by hanging weights and clockwork. Ingenious clocks, puppets which play musical instruments and mechanical calculators stem from that time. The underlying techniques were also applied in experiments to produce artificial speech. In the subsequent section we will review historical mechanical models of speech production from the 18^{th}, 19^{th} and the early 20^{th} century. Current models of speech production are only refinements of these models.

2.2 Ancient mechanical models

2.2.1 The resonators of Kratzenstein

It was in 1779 that the Imperial Academy of St. Petersburg offered its annual prize to the person who could make an apparatus that could render the five vowels A of father, E of they, I of machine, O of note and U of crude. The winner was Christian Gottlieb Kratzenstein, professor of physiology at Halle and later at Copenhagen [3]. He made tube resonators, one for each vowel, the cross-sections of which are depicted in Figure 2.7 [1]. The dimensions of the resonators approximated those of the human vocal tract. They were excited by air blown through a reed, thus mimicking the vocal cords, except the resonator for I, which

[1] See also the web-site www.haskins.yale.edu

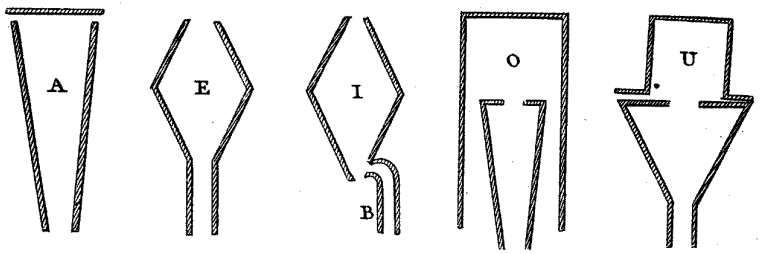

Figure 2.7: Cross-sections of the five resonators of Kratzenstein from 1779 (after Dudley, from the Journal of the Acoustical Society of America [3]).

was excited directly by an uninterrupted airflow like in an organ pipe. It is interesting that Homer Dudley - the author of [3] - also tells the story of the Englishman Robert Willis who showed later (in 1829) that the specific shapes of the resonators of Kratzenstein were not the only solution, and that the different resonances could also be obtained by a single pipe the length of which was tuned to the desired sound. This observation can be seen as an early hint at acoustic tubes, which will be explained later in this chapter.

2.2.2 The speaking machine of von Kempelen

The first speaking machine worthy of the name was constructed by the Hungarian Wolfgang von Kempelen. He was born in Bratislava, now Slovakia, in 1734 and died in Vienna in 1804. He had a high position at the royal court of the Habsburg monarchy and he made the plans of the Schönbrunn fountains in Vienna and the Royal Castle of Buda. Von Kempelen was a recognized organiser and an engineer in heart and soul. He became fascinated by the idea of building a speaking machine and in 1769 he started working on it. In 1791 he published the results of more than two decades of work in a book of 456 pages and 25 illustrations [4]. A summarising description of this book is found in the above mentioned, easier obtainable, paper of Homer Dudley [3]. The book contains studies of speech and the sounds of European languages and it explains the voice, and in particular the functioning of the lungs, trachea, glottis, mouth, nose, tongue, teeth and lips. In illustrations 24 and 25 of his book, as shown in Figure 2.8 and Figure 2.9, von Kempelen shows the top and side views of the final design of his speaking machine.

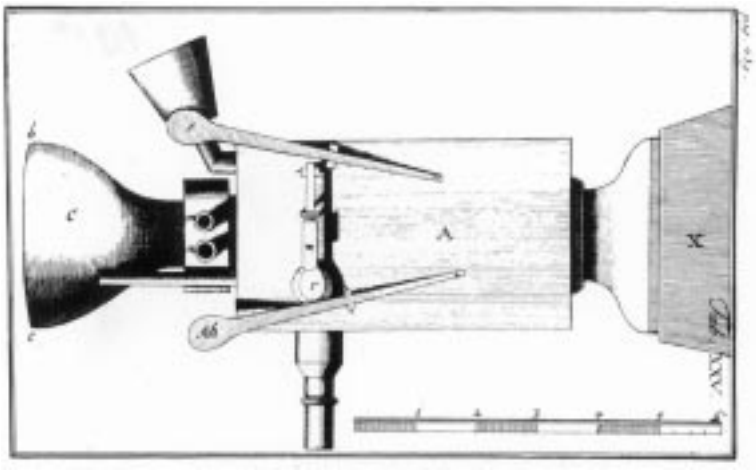

Figure 2.8: Top-view of von Kempelen's speaking machine as shown in his book from 1791.

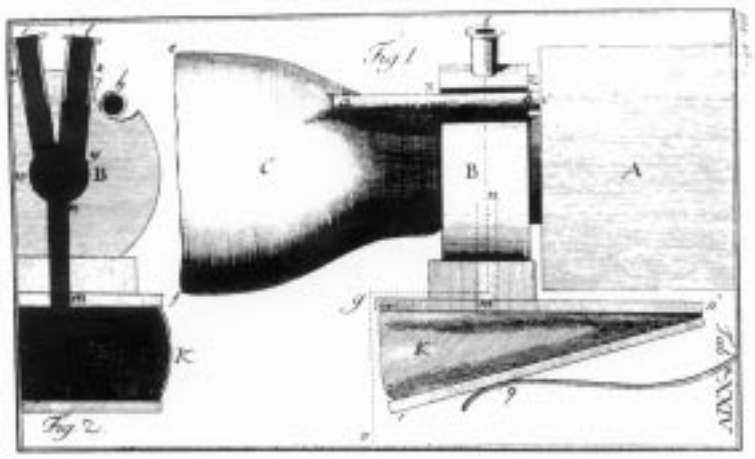

Figure 2.9: Side-view of von Kempelen's machine and cross-section through "B".

The top-view shows a system consisting of a box A which contains a reed from a bagpipe that is excited by airflow from a bellows X. The resulting periodically interrupted airflow leaves the box at the left side and enters a round coupling piece giving access to a bell-shaped resonator C. Different sounds were produced by placing the left hand in different

shapes at different positions in front of the mouth of the resonator. The coupling piece also has two outlets on top, simulating the nostrils (m and n). Nasal sounds could be produced by opening these and at the same time shutting off C. The sound R was produced by using lever r, which lowered a wire until it touched the reed. The lever Sch was used to divert the airflow from the reed to a whistle at one side of the box thus producing the sound "sh" (in German written as sch). A similar lever and a second, differently shaped, whistle at the other side of the box was used to produce the sound "s". The machine was operated by activating the bellows X with ones right arm and elbow leaving the right hand free for controlling the levers and the nostril holes. The ruler shown at the bottom has a length of 6 inches.

The side-view shows some parts that were especially added to produce some plosive sounds like p. These were made by closing C and when sufficient pressure had been built up, by releasing it suddenly. In order to build up enough pressure in a short enough time, it had turned out that a shortcut between the air entering wind-box A and the resonator C with a pipe a-b and an extra bellows K, were both necessary. The cross-section through "B" shows the nostrils in the coupling piece, and the shortcut. The machine could produce Latin, French and Italian words, even complicated ones like "Constantinople" [5]. German words were more difficult because of the complex structure of many of its consonants. von Kempelen could also manage to produce short phrases like "Leopoldus Secundus" [3].

2.2.3 Wheatstone's improved version

The British physicist Sir Charles Wheatstone (1802 - 1875), known for his bridge circuit for comparing resistances, capacitances and inductances, also invented the relay in 1834. The relay became a crucial component (as regenerator, repeater) in long distance telegraphy, as invented by the American Samuel F.B. Morse in 1837 [6]. Professor Wheatstone was also interested in speech and he made a reconstruction of von Kempelen's speaking machine and improved it. He presented it at the Dublin meeting of the British Association for the Advancement of Sciences in August 1835 (see Figure 2.10). The main difference with von Kempelen's speaking machine was the resonator. Wheatstone made a resonator of leather which could be moulded by hand. In this way the desired acoustical characteristic resonances could be better controlled.

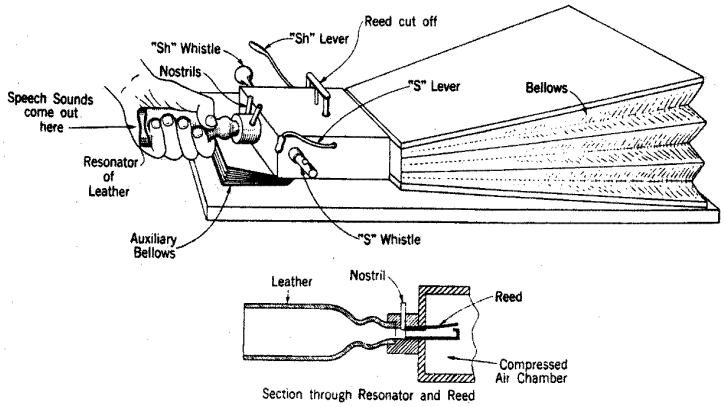

Figure 2.10: Improved version of von Kempelen's speaking machine from 1835 made by Wheatstone (from the Journal of the Acoustical Society of America [3]).

2.2.4 The "Euphonia" of Faber

In 1846, a speaking machine was demonstrated by Professor Josef Faber of Vienna in the Egyptian Hall, Piccadilly, London. He called his machine "Euphonia". The apparatus could be played like an organ by an operator seated at a console, as depicted in Figure 2.11 [3]. Note the artificial mouth with lips and teeth mounted to the upright. This machine could produce whispered speech, normal speech, and simple songs. The big difference with von Kempelen's machine was the attribute of a controllable variable pitch enabling singing.

2.2.5 The speaking device of Bell

Alexander Graham Bell (1847-1922), who later invented the telephone (in 1875) while he was a professor in physiology of the speech organs at the university of Boston, happened to see Wheatstone's speaking machine in Edinburgh, Scotland, during his youth. Impressed by the ingenuity of the design, he started to make a speaking apparatus of his own with the help of his father and his brother. As compared to von Kempelen's machine, the resulting design had more resemblance to the physiologic reality - it consisted of a head of plaster completed with moving parts of different materials imitating the vocal organs, including a sectioned wooden tongue - and it could be actuated by levers from a

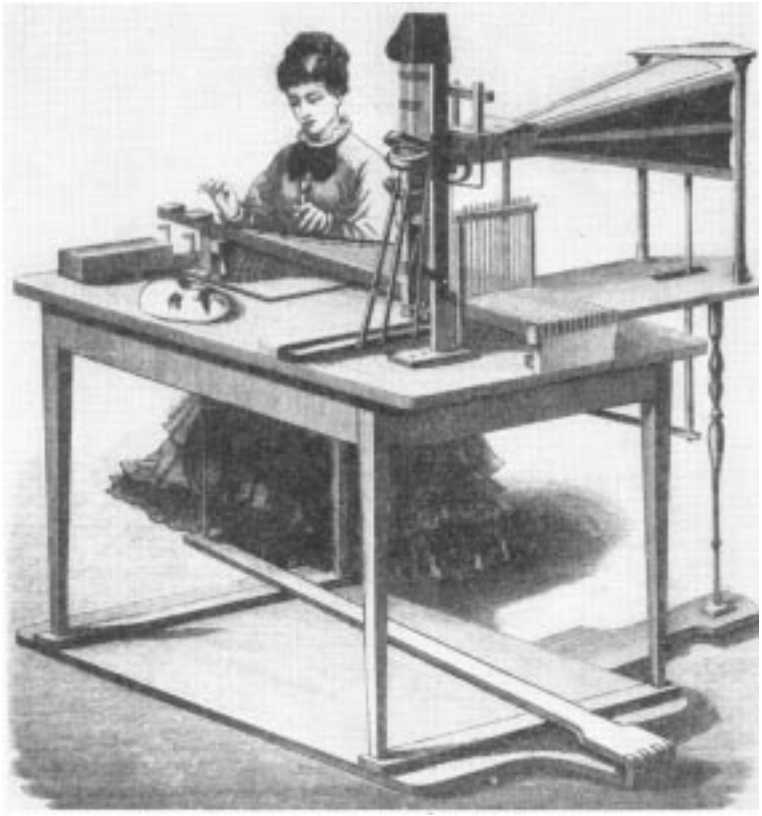

Figure 2.11: The speech organ of Professor Faber from 1846 (from the Journal of the Acoustical Society of America [3]).

kind of keyboard [1]. Bell's device could produce vowels, nasals and a few simple utterances.

2.2.6 The vocal apparatus of Riesz

Another attempt to make a machine speak was undertaken by R.R. Riesz of Bell Telephone Laboratories. In 1930 Riesz reported on an artificial larynx [7] and in 1937 he made a mechanical analogon of the human vocal apparatus the moving parts of which could be controlled by spring-loaded finger keys, like a trumpet, as shown in Figure 2.12. The key numbered 1 controls the lip-opening and key 2 can move the (upper) front teeth. Keys 3, 6 and 7 control the acoustical effect of the tongue and key 8 controls the size of the pharynx. Key 4 operates on the

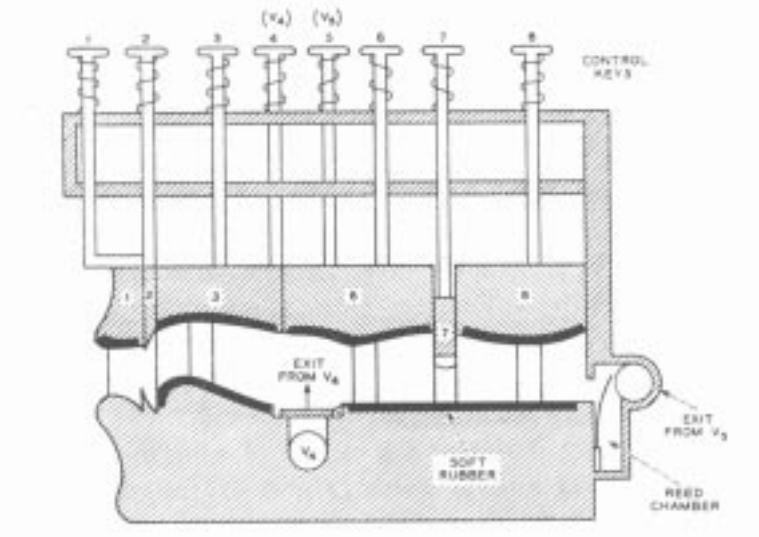

Figure 2.12: Mechanical vocal tract and excitation of Riesz from 1937, operated by finger keys (from the book of Flanagan [8].)

valve V4 through which air is pushed into the mouth cavity to generate unvoiced sounds. The key number 5 controls the airflow from valve V5 into the reed chamber to produce voiced sounds. The reed vibrates at a pitch which can be regulated by the volume velocity of the airflow from V5. Expert operation of the machine could make it say words. One of the words that could be produced particularly well was "cigarette" [8, Section 6.1. on "Mechanical Speaking Machines; Historical Efforts", pp. 205–210].

2.2.7 The tuning forks of von Helmholtz

The models described in the preceding sections are based on more or less lifelike mappings of the physical shapes of the human speech organ. The German scientist Hermann von Helmholtz (1821-1894), one of the founders of the science of acoustics, also experimented with the production of speech sounds, but in quite a different way. He described that in his book which was issued for the first time in 1862, for the sixth time in 1913 and which was translated into English under the title Sensations of Tone [9]. He considered the various (voiced) speech sounds as being built from sinusoidal tones, an insight which he had developed on the

basis of experiments with the aid of his resonators and which had been advocated before by Georg Simon Ohm (1787-1854), the German physicist whose name became forever associated with the unit of electrical resistance [10]. The resonators of von Helmholtz were spheres made of glass or metal which had an opening to the open air on one side and on the other side an opening in a kind of ear-piece fitting the entrance of the ear-channel and which could be pressed onto it. Such a resonator was tuned to a particular frequency by its dimensions. By using different resonators von Helmholtz could analyse sound waves and distinguish tones and overtones.

Von Helmholtz generated various speech vowels by activating a row of accurately harmonically tuned tuning forks and controlling the sound intensity of each individual tuning fork again by an acoustic resonator. Figure 2.13 shows an original drawing of such an instrument which he used around 1862. It consists of eight of these tuning forks and resonators, an oscillator producing a square wave in the form of an interrupted direct current, and two batteries. An individual tone generator is shown in Figure 2.14. It consists of a tuning fork a which is mounted vertically on a base d and which can be excited with the aid of the electro-magnets b. The base rests on rubber tubes e to minimize acoustic coupling with the ground. The sound that is generated by this tuning fork is very weak. It can be amplified with the aid of the tube resonator[2] j which is tuned to the frequency of the tuning fork by means of its dimensions. The distance between the resonator and the tuning fork can be controlled by means of the movable pedestal of the resonator k. The smaller this distance, the higher the amplification. In addition, the sound intensity can be controlled by increasing or decreasing the opening of the resonator with the aid of the lid l on the rod m. This could be done remotely by means of the wire n and the spring p. The wires of all tone generators in the row were connected to a keyboard. By using eight of these generators, thus producing eight harmonic tones, von Helmholtz could fabricate good quality vowels U and O, just by controlling the intensity of the tones. The German vowels E like "a" in day, and I like "i" in machine, could not be produced so well, however. As we know now, many more than eight harmonics are needed for a good reproduction of these sounds. This work can rightly be seen as a predecessor of the sinusoidal model which will be discussed later in this chapter (see Section 2.3.3).

[2]In Figure 2.14 j is found just above the cylinder, although it looks like an i.

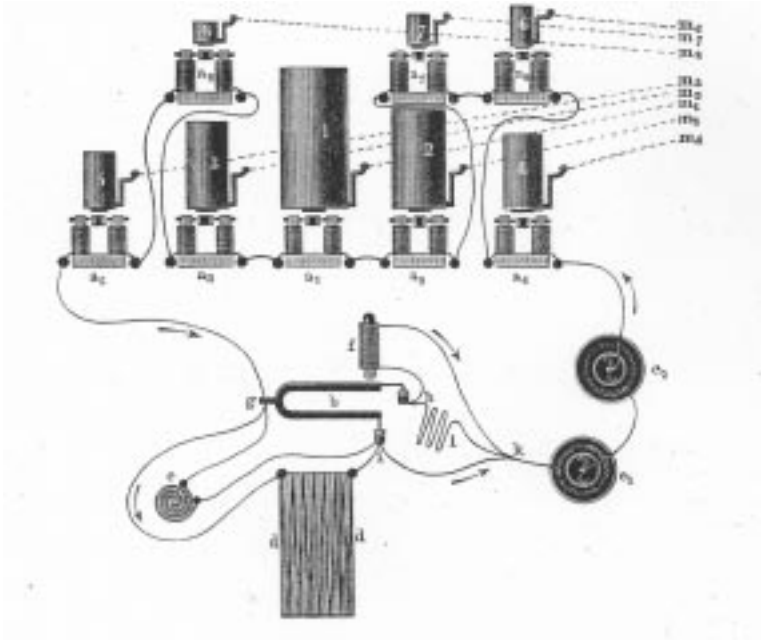

Figure 2.13: Harmonic generator of von Helmholtz capable of imitating vowels, which was already included in the first edition of his book from 1862. The eight tuning-fork based generators (a_1–a_8) are excited by an interrupted direct current from the batteries (e_1 and e_2). The interrupter makes use of a resistor (l), a coil (d), a capacitor (c) and a relay. The relay consists of a tuning fork (b)and an electromagnet (f), and two needles connected to the ends of the tuning fork each of which can make contact with a bath of mercury contained in two small vessels (i and h). (From the book of Helmholtz [9].)

2.3 Early electrical models

In his article in the September 2 issue of the year 1922 of the journal Nature, John Q. Stewart - associated with the American Telephone and Telegraph Company - wrote: "It is well known that Helmholtz succeeded in imitating vowels by combinations of tuning forks, and Miller by combinations of organ pipes. Others, notably Scripture, have constructed apparatus wherein the transient oscillations of air in resonant cavities were excited by series of puffs of air, in close physical imitation of the action of the human vocal organs. It seems hitherto to have been over-

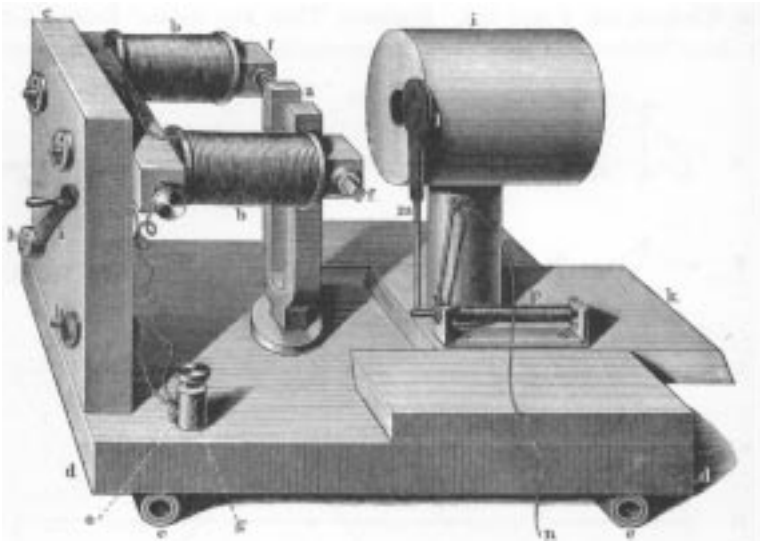

Figure 2.14: The tuning-fork (a) based tone generator of von Helmholtz (from the book of Helmholtz [9]).

looked that a functional copy of the vocal organs can be devised which depends upon the production of audio-frequency oscillations in electrical circuits." [11].

2.3.1 The circuit of Stewart

The circuit that Stewart used to investigate the electronic production of speech sounds is given in Figure 2.15. The circuit consists of an excitation part and a filter part. The excitation makes use of a buzzer or a motor-driven circuit interrupter which together with a battery and a capacitor acts as a relaxation oscillator to produce a sawtooth-like waveform rich of harmonics. The filter part consists of two adjustable tuned circuits to simulate formants, which are coupled to a kind of loudspeaker called "telephone receiver" in Figure 2.15. Stewart could produce several vowels by using periodic interruptions. By the use of non-periodic interruptions he could even produce whispered speech and unvoiced sounds. He also reported that when appropriate arrangements were made to give circuit adjustments in rapid succession, simple words like "mama", "Anna" and "wow-wow" could fairly well be imitated. He also presented a table with used formant frequencies and damping

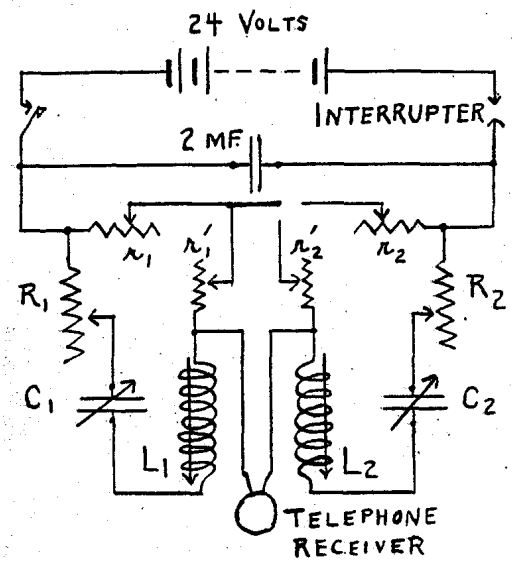

Figure 2.15: The first electronic circuit for the artificial production of speech sounds made by Stewart in 1922 (from Nature [11]).

factors (or equivalently bandwidths) for the various vowels.

Being the first in producing electronic speech, he immediately recognized the problem behind speech synthesis: "The really difficult problem involved in the artificial production of speech-sounds is not the making of a device which shall produce sounds which, in their fundamental physical basis, resemble those of speech, but in the manipulation of the apparatus to imitate the manifold variations in tone which are so important in securing naturalness."

2.3.2 Dudley's Voder

After Stewart's publication, it remained rather quiet for another decade and then suddenly there was Homer William Dudley (1898 - 1981) of Bell Telephone Laboratories in New York, who invented the vocoder - a contraction of the words "voice coder" - in 1935 [12]. This was a great innovation, a real breakthrough. In Dudley's vocoder, speech was analysed into a set of time-varying parameters consisting of pitch, voicing and ten spectral parameters describing the spectral intensity distribution. Solely from these parameters, he could regenerate the speech

Figure 2.16: Mrs. Harper creates speech by operating Dudley's Voder in 1939 (from Bell Laboratories Record [14]). Note the resemblance with the 93 years older console of Faber of Figure 2.11.

signal. The vocoder itself will be further discussed in the next chapter. Here, we will confine ourselves to the speech synthesiser which he later derived from it, in 1937, and which he called the Voder (Voice Operation DEmonsratoR) [13]. The first publication in Bell Laboratories Record [14] explained that it concerned an instrument that could be operated from a console with a keyboard and a pedal, as shown in Figure 2.16, and that a course of training had been developed for its operators. It also mentioned that it would be exhibited by the Bell System at the World's Fair in New York in 1939. In [15] a nice photograph of the exposition booth is given.

The technical details of the Voder are discussed in a paper by Homer Dudley, R.R. Riesz and S.S.A. Watkins [16]. The underlying model

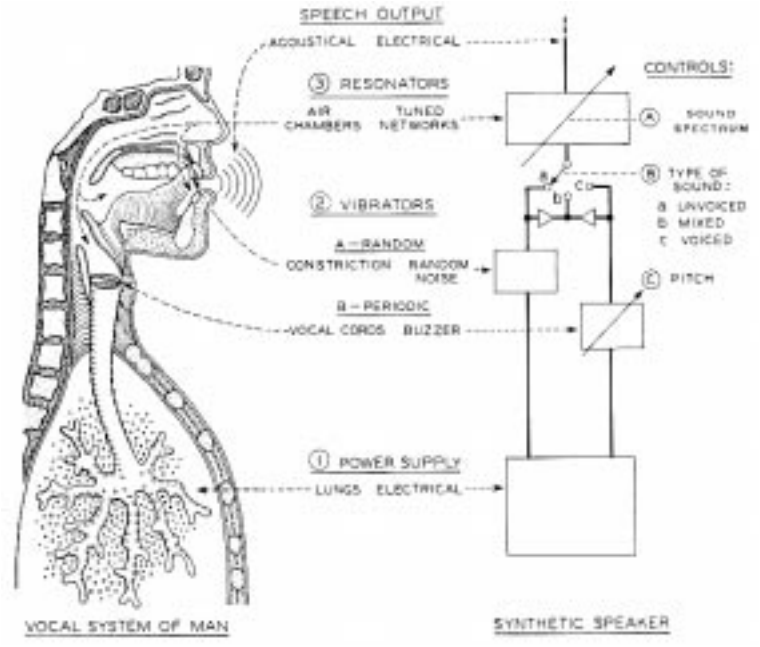

Figure 2.17: Dudley's electronic model of speech production from 1939 (from the Journal of the Franklin Institute [16]).

Dudley developed is depicted in Figure 2.17. The lungs are modelled as a power supply. The vocal cords are modelled as a relaxation oscillator ("buzzer"), the pitch of which is controllable. For unvoiced sounds the model provides for a random noise source. The excitation signal "B" consists of a) unvoiced excitation from the noise source or b) mixed excitation or c) voiced excitation from the buzzer. The mixed excitation consists of a linear combination of a and c. The vocal tract is modelled as a bank of (parallel) tuned networks with amplitude control in each branch. Although the paper does not give any details about the filter bank, the patents do. Both patents [12] and [13] suggest a logarithmic distribution of the bandwidths of the ten bandpass filters, for instance starting with 0 - 225 Hz and 225 - 450 Hz for the first two filters and then increasing bandwidths for the next filters up to 5400 - 7500 Hz for the last filter. Patent [12] suggests an attenuation of about 6 dB at the separating points between two bands, 20 dB attenuation at the middle of the next band and 40 dB at the middle of the second band.

The Voder's operating console, see Figure 2.18, had ten finger keys

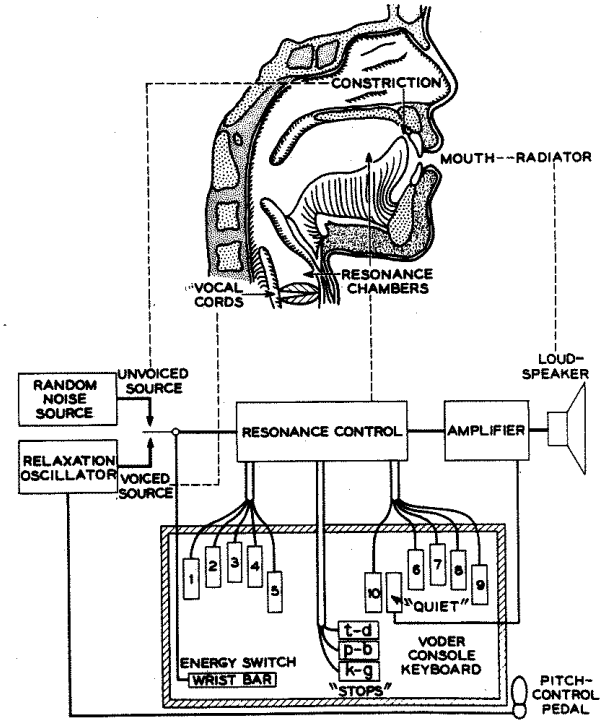

Figure 2.18: Dudley's Voder and its operating console functions (from the Journal of the Franklin Institute [16]).

for the amplitude control of the filter-bank outputs ("resonance control"), another "quiet" key for inserting a 20 dB attenuation, especially for the natural production of unvoiced sounds, three finger keys for controlling stops, a bar operated by the wrist controlling the voicing switch and a pedal for pitch control. The stop-keys facilitated the production of plosive and stop consonants. The t-d key was used for the production of t using unvoiced excitation or for d using voiced excitation. The p-b and the k-g keys served a similar purpose.

In many ways the Voder could perform operations similar to those performed by the human speech organs. The Voder could not only produce speech and chant over a wide range from bass to soprano voices but it could also produce non-speech sounds, such as imitations of a cow, a pig, a woodpecker, various insects, an air-plane, and a steam-train. It took about a year for an operator to acquire the ability to

produce speech sounds with sufficient naturalness and intelligibility.

2.3.3 Dudley's "carrier" model

In 1940, Dudley devoted a publication to the "carrier nature" of speech [17]. In this article he explained new insights on the basis of carrier circuit techniques, a frequency division multiplexing (FDM) technique widely applied in telephone circuits in those days. The most interesting part of this article is the appendix in which he gave mathematical relations of the carrier nature. He defined a carrier C by a sum of n harmonic sinusoids according to

$$C = \sum_{k=1}^{n} A_k \cos(kPt + \varphi_k), \tag{2.1}$$

in which A_k stands for the amplitude, φ_k for the phase and kP (P for pitch) for the frequency of a sinusoid. Subsequently, he described that the amplitude and pitch parameters can be modulated to make them time-varying so as to imitate voiced speech. Dudley also proposed a model for unvoiced sounds, based on the same formula with $P \rightarrow 0$ and $n \rightarrow \infty$.

This work covers nothing less than a sinusoidal model, as already advocated by Ohm and von Helmholtz, but it got no subsequent follow-up, not even by Dudley himself. Dudley looked back on his achievements in a paper presented at the Seventh Annual Convention of the Audio Engineering Society (AES), in October 1955, in which the "carrier nature" was hardly mentioned anymore [18]. In the same paper Dudley recommended the electrical vocal tract of Dunn which provided much better vowel quality than the Voder, as demonstrated at the same AES Convention.

In the 1980s the sinusoidal model will be reintroduced again, with success, as mentioned in Chapter 5.

2.3.4 Dunn's acoustic model of the vocal tract

Dunn modelled the pharynx, the constriction of the hump of the tongue, the mouth cavity and the constriction due to the lips as a concatenation of four cylindrical tube-sections, as shown in Figure 2.19 [19]. The cross-sectional areas of the sections are indicated by $\mathcal{A}$ and the length of the sections by l. The piston at the left side of the composite tube is supposed to generate a plane wave with a volume velocity (the volume

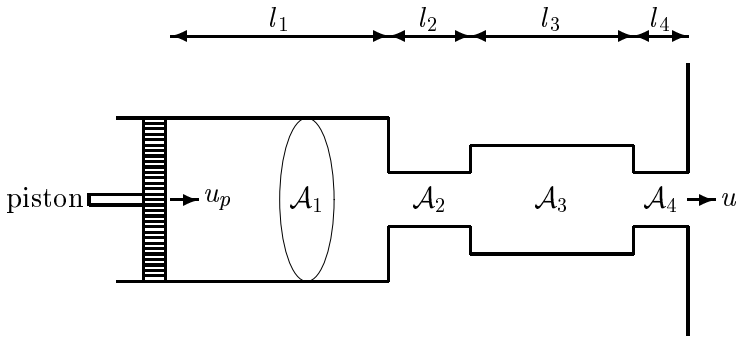

Figure 2.19: Dunn's model of the vocal tract consisting of four cylindrical tube sections with different cross-sectional areas ($\mathcal{A}$) and lengths (l), 1950.

of gas that passes a location per unit of time) u_p corresponding to the glottal excitation waveform. At the other end the tube is terminated by a plane baffle modelling the radiation from the mouth in the head, and the output volume velocity is indicated by u. Dunn succeeded in developing a method to determine the dimensions of this tube on the basis of the dimensions of the human vocal tract obtained from x-ray photographs, and to fine-tune three artificial formant frequencies of the composite tube to the original formants, for most vowels. As an example, the dimensions he found for the vowel "o" as in "lost", are given in Table 2.1. Earlier attempts following similar approaches, such as those

index	1	2	3	4	
length l	3	4	8	1	cm
area $\mathcal{A}$	2.8	0.75	10.9	6.3	cm^2
formant	640	930	2580		Hz

Table 2.1: Dimensions of Dunn's tube model for the vowel "o" as in "lost".

of Paget [20] and Crandall [21], were not fully successful in linking (only two) formant frequencies to the vocal tract dimensions. They used double Helmholtz resonators of plasticine and cardboard for modelling the throat and mouth cavities specified by their volumes and the conductivities of the coupling (an open connection) between the cavities and the

coupling to the open air.

The main features of Dunn's method were that he used additional cavities for the couplings and that he could calculate the transfer function $U(\omega)/U_p(\omega)$ of such a composite tube from its physical dimensions (U stands for the Fourier transform of u). He used the analogy of a single tube section to a section of an electrical transmission line. This was common practice already in those days as evident, among other documents, from the textbook of Philip Morse from 1948 [22, Chapter VI: "Plane waves of sound"]. The application of this technique to the vocal tract modelling problem, however, was the inception of a model that we still use today. Before we continue our review of speech production models with a more general acoustic tube model, we will briefly revisit the underlying theory of this analogy. Whereas the historical models have been discussed in a narrative style up till now, the acoustic tube models (and later also the technique of linear prediction as applied to speech), are subject to more mathematical analysis in order to create a sound basis for the treatise of Chapter 4.

2.3.5 Electrical model of a uniform acoustic tube

Let us consider a thin cross-sectional slice of a tube-section with area $\mathcal{A}$. A sound wave travelling through the slice is assumed to be a plane wave so that the local pressure variation $p(x,t)$ and volume velocity variation $u(x,t)$ are only functions of the longitudinal dimension x and the time t, and not of the dimension perpendicular to x. This is usually the case if the diameter of the tube is small with respect to the wavelength of the sound. For the vocal tract this condition is adequately met for frequencies up to a few kHz. Furthermore, our purpose is sufficiently met by confining ourselves to the lossless case. Acoustic losses in the tube are caused by viscous resistance of the gas in the tube, in our case air, and by energy loss via the walls of the tube. The latter cause can be prevented by assuming a rigid tube with negligible heat conductance. Under the lossless conditions the relation between $p(x,t)$ and $u(x,t)$ is given by the wave equations

$$-\frac{\delta p(x,t)}{\delta x} = \frac{\varrho}{\mathcal{A}} \frac{\delta u(x,t)}{\delta t} \tag{2.2}$$

and

$$-\frac{\delta u(x,t)}{\delta x} = \frac{\mathcal{A}}{\eta P} \frac{\delta p(x,t)}{\delta t} \ , \tag{2.3}$$

in which ϱ stands for the density of air, η for the adiabatic constant ($\eta = 1.4$ for air) and $P = P_0 + p(x,t)$ for pressure with P_0 representing the (atmospheric) pressure in quiet. The first equation arises from the application of Newton's law of motion to the cross-sectional slice and the second from the application of the adiabatic gas law to the slice. They are associated with the inertia of the mass and the compliance or compressibility of the slice of air, respectively. In Appendix A, a derivation of these equations is given.

The general solution of the acoustical wave Equations 2.2 and 2.3 is given by

$$p(x,t) = \frac{\varrho c}{\mathcal{A}} \left\{ u^+ \left(t - \frac{x}{c} \right) + u^- \left(t + \frac{x}{c} \right) \right\} \tag{2.4}$$

and

$$u(x,t) = u^+ \left(t - \frac{x}{c} \right) - u^- \left(t + \frac{x}{c} \right) \tag{2.5}$$

with

$$c = \sqrt{\eta \frac{P}{\varrho}}, \tag{2.6}$$

in which $u^+(t - x/c)$ stands for the incident wave travelling in the positive x-direction and $u^-(t + x/c)$ for the reflected wave, if any, travelling in the negative x-direction. The given solution can be verified by substitution of Equations 2.4, 2.5 and 2.6 into 2.2 and 2.3. In doing so, it is convenient to utilize the mathematical fact that

$$\frac{\delta f\left(t \pm \frac{x}{c}\right)}{\delta t} = \pm c \frac{\delta f\left(t \pm \frac{x}{c}\right)}{\delta x} \tag{2.7}$$

for any differentiable function $f(t \pm \frac{x}{c})$.

On inspection of the properties of, for instance, the incident wave one can imagine at a certain moment t_0 a waveform as a function of x. Because of the "travelling" property, this waveform will be displaced over a distance Δx at a time Δt later, so the propagation speed is defined by $\Delta x / \Delta t$. The shape of the waveform itself is still the same, however, so that:

$$u^+ \left(t_0 - \frac{x}{c} \right) = u^+ \left(t_0 + \Delta t - \frac{x + \Delta x}{c} \right). \tag{2.8}$$

By equating the indices it follows that c must be the propagation speed of the wave.

Now, let us consider an electrical transmission line consisting of two parallel conductors. If we assume an electrical signal source on one end

of the line, there will be signal loss along the line towards the other end, which we will assume to be the positive x-direction, again. The signal loss is caused by series inductance and resistance of the two conductors and by leakage between the two conductors due to shunt capacitance and shunt conductance. If we confine ourselves to the lossless case again, and we assume a uniform distribution of the inductance and capacitance over the line, then a thin cross-sectional slice of the line can be represented by the circuit of Figure 2.20, in which $\mathcal{L}$ stands for the inductance per unit of length and $\mathcal{C}$ for the capacitance per unit of length. The equations

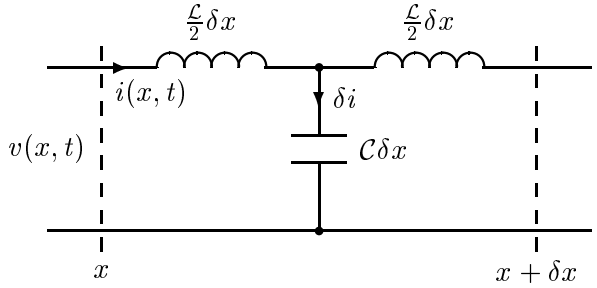

Figure 2.20: Equivalent circuit for a cross-sectional slice of a lossless electrical transmission line with uniformly distributed inductance $\mathcal{L}$ and capacitance $\mathcal{C}$.

which directly follow from the well known relations between the voltages over and the currents through the inductance and the capacitance are:

$$-\frac{\delta v(x,t)}{\delta x} = \mathcal{L}\frac{\delta i(x,t)}{\delta t} \tag{2.9}$$

and

$$-\frac{\delta i(x,t)}{\delta x} = \mathcal{C}\frac{\delta v(x,t)}{\delta t}\,, \tag{2.10}$$

where v and i represent voltage and current, respectively.

If we compare these electrical wave equations to the acoustical ones of Equations 2.2 and 2.3, we see that they are analogous. We infer that pressure p is analogous to voltage v, volume velocity u to current i, and that the "acoustic inductance" and "acoustic capacitance" are respectively represented by

$$\mathcal{L} = \frac{\varrho}{\mathcal{A}} \tag{2.11}$$

and

$$\mathcal{C} = \frac{\mathcal{A}}{\eta P}. \tag{2.12}$$

Due to the analogy with Equations 2.2 and 2.3, the solution of the electrical wave equations, Equations 2.9 and 2.10, must be

$$v(x,t) = Z_0 \left\{ i^+ \left(t - \frac{x}{c} \right) + i^- \left(t + \frac{x}{c} \right) \right\} \tag{2.13}$$

and

$$i(x,t) = i^+ \left(t - \frac{x}{c} \right) - i^- \left(t + \frac{x}{c} \right) \tag{2.14}$$

with

$$Z_0 = \sqrt{\frac{\mathcal{L}}{\mathcal{C}}} \tag{2.15}$$

which represents the characteristic impedance of the transmission line, and

$$c = \sqrt{\frac{1}{\mathcal{L}\mathcal{C}}}\,. \tag{2.16}$$

Before we will determine the transfer function of a uniform acoustic tube of length l on the basis of a transmission line of the same length, we will consider the case of infinite l. In the tube, the reflected wave comes back from the output where the physical dimensions change and no reflections take place in the uniform tube itself (see Section 2.4.1). So, in an infinitely long acoustic tube which is excited at one end, the reflected wave never arrives at any location, so that $u^-(t + x/c) = 0$. From Equations 2.13 and 2.14 we see that under this condition the ratio of $v(x,t)$ and $i(x,t)$, the impedance, is always Z_0 at any location x.

For finite l, the reflected wave will arrive at the input, reflected back again and so on until a steady state situation appears. This typically occurs for periodic excitations. For this purpose we assume complex exponential signals

$$i^+ \left(t - \frac{x}{c} \right) = I^+ e^{j\omega\left(t - \frac{x}{c}\right)} \tag{2.17}$$

and

$$i^- \left(t + \frac{x}{c} \right) = I^- e^{j\omega\left(t + \frac{x}{c}\right)}, \tag{2.18}$$

and the constants I^+ and I^- are determined by the terminal conditions

$$i(0,t) = \left(I^+ - I^-\right) e^{j\omega t} \tag{2.19}$$

and

$$i(l,t) = \left(I^{+}e^{-j\omega\frac{l}{c}} - I^{-}e^{j\omega\frac{l}{c}}\right)e^{j\omega t} \ . \tag{2.20}$$

Combining Equations 2.19 and 2.20 with 2.17 and 2.18 and substitution of the result into 2.13 and evaluating it for $x = 0$ and for $x = l$ finally yields

$$v(0,t) = -jZ_0\left\{i(0,t)\cot\left(\omega\frac{l}{c}\right) - i(l,t)\csc\left(\omega\frac{l}{c}\right)\right\} \tag{2.21}$$

and

$$v(l,t) = -jZ_0\left\{i(0,t)\csc\left(\omega\frac{l}{c}\right) - i(l,t)\cot\left(\omega\frac{l}{c}\right)\right\}. \tag{2.22}$$

These relations are represented by the T-network of Figure 2.21, which consequently represents the transfer function of the uniform acous-

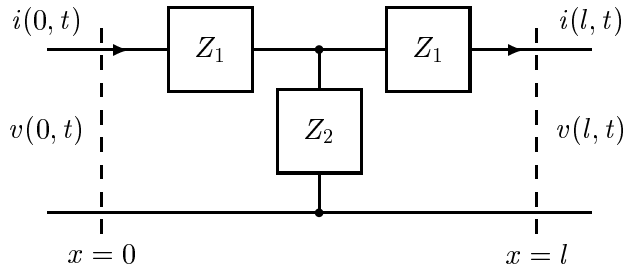

Figure 2.21: Electrical T-network model of a uniform acoustic tube of length l.

tic tube of length l, with

$$Z_1 = jZ_0\frac{1-\cos\omega\frac{l}{c}}{\sin\omega\frac{l}{c}} = jZ_0\tan\left(\omega\frac{l}{2c}\right) \tag{2.23}$$

and

$$Z_2 = \frac{Z_0}{j\sin\left(\omega\frac{l}{c}\right)} = -jZ_0\csc(\omega\frac{l}{c}) \ . \tag{2.24}$$

It is very instructive to determine the transfer function of the tube with the following terminal conditions. At one end, the tube is acoustically activated by a closely fitting, but moving, piston as shown in Figure 2.19 and the other end of the tube is provided with the plane baffle. The

piston is assumed to generate a plane incident wave and it represents a high impedance which reflects all waves that propagate into the negative x-direction. The closed end of the tube can be interpreted as a preceding tube section with a value of $\mathcal{A}$ approaching zero which gives rise to a very high characteristic impedance Z_0 of this virtual section. Therefore, we assume that in the electrical system we can represent the piston as a current source. At the other end of the system the baffle represents a very low impedance and therefore we model it in the electrical system as a short circuit. In first approximation the baffle may be interpreted as a next section with a large value of $\mathcal{A}$. According to Equations 2.11 and 2.12 this gives rise to a small value of $\mathcal{L}$ and a large value of $\mathcal{C}$. According to Equation 2.15 this results in a small value for Z_0, and according to Equations 2.23 and 2.24 this also results in small values of Z_1 and Z_2.

Thus, the transfer function of the acoustic system is approximated by the transfer function of the electrical network of Figure 2.21 with a current source $i(0,t)$ at the input and a short-circuit at the output with $v(l,t) = 0$. Some calculation then yields

$$\frac{I(l,\omega)}{I(0,\omega)} = \frac{Z_2}{Z_1 + Z_2} = \frac{1}{\cos \omega\tau} \,, \tag{2.25}$$

where I stands for the Fourier transform of i and $\tau = l/c$ for the travelling time of the wave over the distance l. This transfer function shows resonances at the frequencies

$$\omega_k = \frac{\pi}{2\tau}(1 + 2k) \tag{2.26}$$

for all integer k and the periodic character of the circular function is manifested by the higher resonance modes. Note that the absolute value of the transfer function has a period of $1/(2\tau)$ Hz. Since we are discussing the lossless case the resonances are represented by poles at those real frequencies. In practice there are always some losses and the resonances will be less sharp.

2.3.6 Dunn's electrical vocal tract model

The eventual electrical model Dunn derived in this way contained 25 equal T-network sections and a few additional inductances [19]. Each T-network represented a uniform acoustic tube with a length of 0.5 cm

and a cross-sectional area of 6 cm^2. In this case $\omega l << c$, so that the impedances Z_1 and Z_2 are very well approximated by respectively

$$Z_1 \simeq j\omega \frac{\mathcal{L}}{2} l \tag{2.27}$$

and

$$Z_2 \simeq \frac{1}{j\omega \mathcal{C} l}, \tag{2.28}$$

and that each T-network can be realized by two equal series inductances with the midpoint connected to a shunt capacitance. The model was provided with the possibility to insert a variable series inductance, representing the tongue hump, at a variable place in this cascade. The tongue hump represents a constriction with a relatively small cross-sectional area, so that according to Equations 2.11 and 2.12 the acoustic inductance becomes large and the acoustic capacitance low. The cascade was appropriately terminated by a variable inductance representing the constriction by the lips, in series with a fixed inductance representing the acoustic radiation from the mouth (see Section 2.5).

All English vowels could be produced by the apparatus, although for some vowels a clear "nasal" character could be perceived. The system had the severe limitation that only three variables could be varied, i.e. the inductance representing the (narrow) passage above the tongue hump, the place of this inductance in the cascade and the inductance representing the area of the mouth opening (the lip constriction). These parameters actually represent the acoustic system of Figure 2.19 with only limited control over the cross-sectional areas and the section lengths.

2.3.7 Terminal analog models

Another contemporary development in modelling the vocal tract was based on the notion that it is not necessarily a model according to the physical dimensions of the vocal tract that is needed, but that any electrical network with an appropriate transfer function is also a good model. The transfer function is defined between the "terminals" of the network and as long as the transfer function of the electrical network is analogous to the transfer function of the acoustical vocal tract, the structure of the network itself does not matter. However, the number of parameters and the way in which these parameters are controlling the time-varying transfer function of the model, do matter. Because of this,

terminal analog models have been exclusively considered as formant-oriented networks, just as the model of Stewart (discussed in Section 2.3.1). The networks consist of resonators with a minimum number of parameters for resonance frequency and bandwidth or magnitude of the resonance. Perhaps the most representative discussion on this topic is in the paper of Flanagan on terminal-analog speech synthesisers from 1957 [25], where parallel and cascade connections of simple electrical resonators form the components of the synthesisers. Later, Flanagan devoted a subsection of his book [8] to terminal-analog synthesisers.

2.4 Composite acoustic tube model

The acoustic tube model of the vocal tract which has turned out to be very useful consists of a series of cylindrical tube sections which all have different diameters but equal lengths. Figure 2.22 shows an example of such a composite tube. The total length is supposed to match the length of the vocal tract which is about 17 cm. This is different as compared to Dunn's tube which consisted of four sections having different diameters and different lengths. The first publication which suggests such a construction with exclusively equal-length sections is the paper by Stevens, Kasowski and Fant from 1953 [23]. In that paper each section represents 0.5 cm length of the vocal tract and cross-sectional areas can vary from 0.17 to 17 cm^2. In the described experimental set-up the areas were controlled by switch settings. The authors reported that "...it has been possible to synthesise all the English vowels. The quality of the vowels is judged by experienced listeners to be good." In this section we derive the transfer function of such an acoustic tube and we will discuss several alternative network representations, i.e. ladder, lattice and direct-form networks. In addition, networks having the inverse transfer function will be determined. The acoustic tube turns out to be directly linked to linear prediction, which is the subject of Section 2.5. The counter-intuitive backward numbering of the sections in Figure 2.22 supports this direct link and complies with linear prediction conventions. Most of the relevant literature deals with both subjects jointly. Therefore, this literature will mainly be discussed in Section 2.5.

2.4.1 Reflection coefficients and the ladder network

In Figure 2.22 several parameters are indicated which are of relevance for our purpose. The composite tube is supposed to have $M+1$ sections with

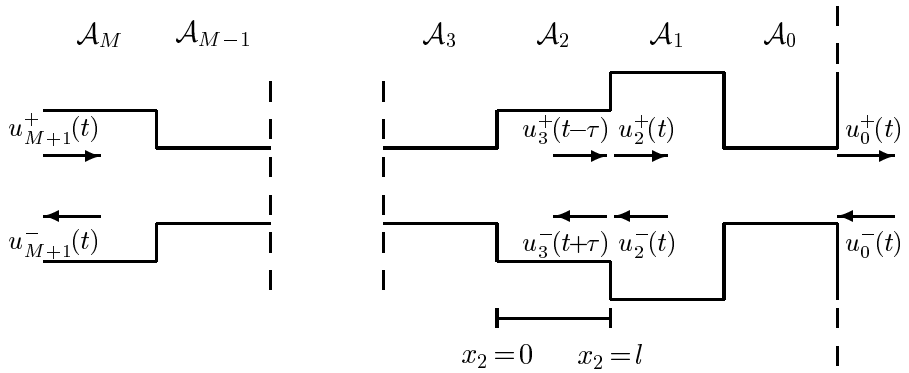

Figure 2.22: Composite acoustic tube with $M+1$ equal-length sections having different cross-sectional areas $\mathcal{A}_m$, 1950.

cross-sectional areas $\mathcal{A}_m$, $m = 0, 1, ...M$, and a longitudinal parameter x_m ranging from 0 to l, for each section m. Furthermore, we assume that the diameters of the sections in the tube are small with respect to the wavelength so that plane waves can be assumed throughout the composite tube, and that there are no energy losses. From the previous section we know that in the case of plane waves the behaviour of each tube-section can be described in terms of the incident wave u^+ and the reflected wave u^-. In Figure 2.22 these waves are indicated at the input of the composite tube, at the junction of two sections (sections 1 and 2) and at the output of the composite tube. We will first deal with the situation around the junction and then return to the input and output conditions.

The incident wave in section 2 at $x_2 = 0$ is denoted as the output incident wave $u_3^+(t)$ of section 3. At $x_2 = l$ we will find some time later the same wave that has travelled over a distance l. The travelling time τ amounts to

$$\tau = \frac{l}{c}\,, \tag{2.29}$$

so we can denote the incident wave at $x_2 = l$ by $u_3^+(t-\tau)$. In the same way, the reflected wave towards section 3, $u_3^-(t)$ at $x_2 = 0$, travels in the negative x-direction and a time τ earlier it was at the location $x_2 = l$, so that it can be denoted by $u_3^-(t+\tau)$ at this location. At the very same location, but now for $x_1 = 0$ in section 1, the incident and reflected waves are consequently given by $u_2^+(t)$ and $u_2^-(t)$, respectively. Because

physics does not allow discontinuities in volume velocity at the junction location, we can postulate the condition

$$u_{m+1}(l,t) = u_m(0,t) \ , \tag{2.30}$$

and the same holds for the pressure, so

$$p_{m+1}(l,t) = p_m(0,t) \ . \tag{2.31}$$

By substitution of Equations 2.4 and 2.5 and the above denotations of the respective incident and reflected waves into these junction conditions and by writing the waves travelling away from the junction explicitly as a function of the waves travelling towards the junction, we obtain

$$u_m^+(t) = (1 + r_m)u_{m+1}^+(t - \tau) + r_m u_m^-(t) \tag{2.32}$$

and

$$u_{m+1}^-(t + \tau) = -r_m u_{m+1}^+(t - \tau) + (1 - r_m)u_m^-(t) \ , \tag{2.33}$$

where r_m is given by

$$r_m = \frac{\mathcal{A}_{m-1} - \mathcal{A}_m}{\mathcal{A}_{m-1} + \mathcal{A}_m} \ . \tag{2.34}$$

The coefficients r_m are the so-called reflection coefficients[3]. For non-negative areas $\mathcal{A}_m$ they are evidently bounded by unity:

$$-1 \leq r_m \leq 1 \ . \tag{2.35}$$

The function of the Equations 2.32 and 2.33 is represented by the signal-flow graph of Figure 2.23. It is known as the Kelly-Lochbaum model [24]. The delays τ represent the travelling time through the tube-section m and the ladder-shaped network represents the junction with the next tube-section $m - 1$. It illustrates very well the nature of the reflection coefficients. A wave travelling towards the junction is partly reflected and the complement is passed on. Because the reflection can occur in anti-phase, the complement can exceed the original wave in amplitude so that the sum of the two components always yields the original wave.

By concatenation of $M + 1$ of these ladder-network sections, each with its own reflection coefficient, a signal-flow graph of the composite tube is obtained.

[3] In the literature the reflection coefficient is sometimes defined with the opposite sign.

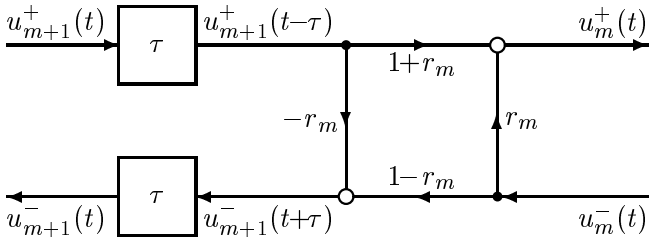

Figure 2.23: Ladder-network section.

The excitation of the tube from a virtual section $M+1$ is supposed to cause a plane incident wave $u^+_{M+1}(t)$ in section M and the reflected wave $u^-_{M+1}(t)$ is supposed to leave the tube un-hindered without any back-reflection into the tube again, in agreement with $r_{M+1} = 0$. The output of the composite tube is terminated by a plane baffle which can be interpreted as a virtual tube section with a very large cross-sectional area so that $r_0 = 1$. In addition, it is supposed that no reflected wave exists so that $u^-_0(t) = 0$.

The transfer function of this continuous-time composite tube, $H_c(\omega)$, is defined as

$$H_c(\omega) = \left.\frac{U^+_0(\omega)}{U^+_{M+1}(\omega)}\right|_{U^-_0(\omega)=0}, \tag{2.36}$$

where $U(\omega)$ stands for the Fourier transform of $u(t)$.

2.4.2 Discrete-time models and lattice networks

A discrete-time model is obtained by replacing the delay τ by the discrete-time delay operator z^{-1}, assuming that the sampling frequency f_s equals $f_s = 1/\tau$. It is easy to show that the transfer functions of both the discrete-time and the continuous-time versions are equal. This duality is explained in Appendix B. The sampling frequency, however, can under certain conditions be halved to $f_s = 1/(2\tau)$, as already suggested by Equation 2.26, which gives rise to a more efficient discrete-time model. In addition, the four multiplications of each ladder section are reduced

to two multiplications in a lattice section. These items will be elaborated next.

It is straightforward to verify that both networks of Figure 2.24 are functionally equivalent to the original ladder section m of Figure 2.23 and that they both satisfy Equations 2.32 and 2.33. Notice that the de-

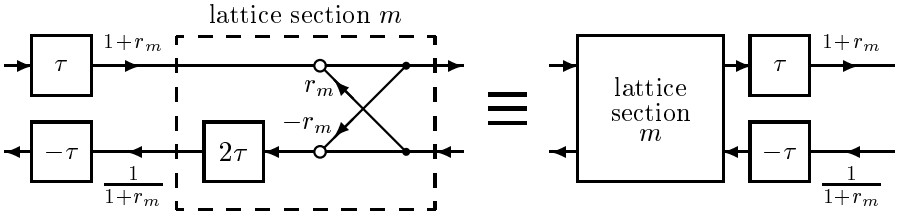

Figure 2.24: Functionally equivalent sections containing lattice sections.

lay pair $(\tau, -\tau)$ can be propagated through the network, just as the gain pair $(1 + r_m, 1/(1 + r_m))$. By applying these properties to the composite tube we find the equivalent network for a chain of $M + 1$ ladder sections as shown in Figure 2.25. Notice that the reflected wave $u^-_{M+1}(t)$ at the

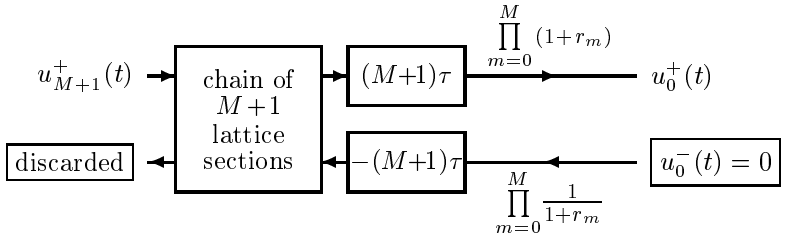

Figure 2.25: Chain of lattice sections.

input will be discarded and that the reflected wave at the output is set to zero, $u^-_0(t) = 0$, as mentioned before.

For the discrete-time model the delays 2τ in the lattice sections are replaced by z^{-1}, assuming a sampling frequency of $f_s = 1/(2\tau)$. The accumulated delay and gain pairs will be discarded in the discrete-time model because their presence does not serve any particular purpose while unnecessary delays, possibly incorporating problematic delays of half a

sampling period, are avoided. The resulting chain of lattice sections still does not have the desired form. As we will see in Chapter 4, the GSM coder contains a slightly different, but functionally equivalent, lattice network. This is based on the identity shown in Figure 2.26. The

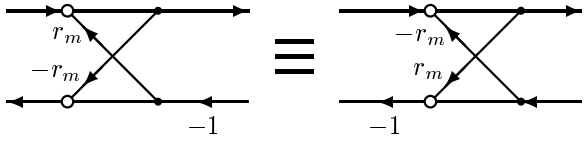

Figure 2.26: Two functionally equivalent networks showing negation propagation.

identity can easily be verified by comparing the network equations analogous to Equations 2.32 and 2.33. Notice that the gain factor -1 in the bottom branch can be propagated through the network while the reflection coefficients change signs. This is applied to the chain in the following way. The right-hand side of the last lattice section in the chain (section 0) has a zero input (bottom branch) and a connection from the output (top branch) to the bottom branch with a weight $-r_0 = -1$. This factor -1 is propagated through the chain of $M + 1$ sections all the way until the discarded bottom output of section M. So, the functionally equivalent network is the same, but all reflection coefficients have changed signs. Thus, the discrete-time model consists of a chain of lattice networks as shown in Figure 2.27. Because it models a part

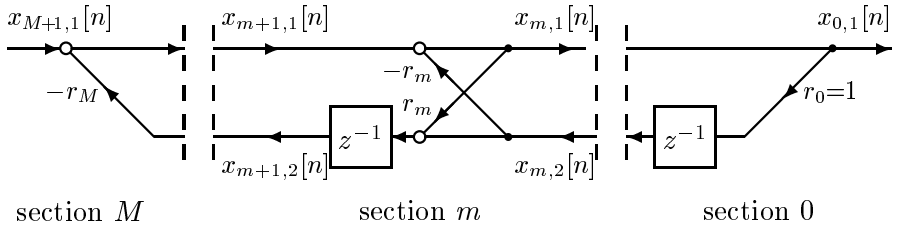

Figure 2.27: Lattice synthesis filter.

of the speech synthesis process, it is referred to as the lattice synthesis filter. In the figure, different signal notations are used to emphasise that

no one-to-one relation to the acoustic volume velocity waves u exists anymore, because the signals are in the discrete-time domain and they are associated with different delays and signal levels.

2.4.3 Transfer function and inverse filter

In the determination of the transfer function $H(z)$ of the lattice synthesis filter the z-transform of $x_{m,j}[n]$ will be denoted by $X_{m,j}(z)$, and the following procedure is applied. Suppose we know the output $X_{0,1}(z)$, then we also know $X_{1,1}(z)$ and $X_{1,2}(z)$ and so on till we find $X_{M,1}(z)$ and $X_{M,2}(z)$ and finally $X_{M+1,1}(z)$. The transfer function is then given by

$$H(z) = \frac{X_{0,1}(z)}{X_{M+1,1}(z)} \,. \tag{2.37}$$

In doing so, we represent section m by

$$\begin{bmatrix} 1 & r_m \\ r_m z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} X_{m,1}(z) \\ X_{m,2}(z) \end{bmatrix} = \begin{bmatrix} X_{m+1,1}(z) \\ X_{m+1,2}(z) \end{bmatrix} \,. \tag{2.38}$$

For the first step, describing section 0, we can write

$$\begin{bmatrix} 1 & r_0 \\ r_0 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} X_{0,1}(z) \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & r_0 \\ r_0 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} X_{0,1} = \begin{bmatrix} X_{1,1}(z) \\ X_{1,2}(z) \end{bmatrix} . \tag{2.39}$$

For section 1 we write

$$\begin{bmatrix} 1 & r_1 \\ r_1 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} X_{1,1}(z) \\ X_{1,2}(z) \end{bmatrix} = \begin{bmatrix} X_{2,1}(z) \\ X_{2,2}(z) \end{bmatrix} \,, \tag{2.40}$$

and substitution of Equation 2.39 into Equation 2.40 yields

$$\begin{bmatrix} 1 & r_1 \\ r_1 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 & r_0 \\ r_0 z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} X_{0,1}(z) = \begin{bmatrix} X_{2,1}(z) \\ X_{2,2}(z) \end{bmatrix} \,. \tag{2.41}$$

Continuing this procedure and representing section M by

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & r_M \\ r_M z^{-1} & z^{-1} \end{bmatrix} \begin{bmatrix} X_{M,1}(z) \\ X_{M,2}(z) \end{bmatrix} = X_{M+1,1}(z) \,, \tag{2.42}$$

finally yields a polynomial $A(z)$ such that

$$A(z) X_{0,1}(z) = X_{M+1,1}(z) \tag{2.43}$$

where

$$A(z) = \begin{bmatrix} 1 & 0 \end{bmatrix} \left\{ \prod_{i=0}^{M} \begin{bmatrix} 1 & r_{M-i} \\ r_{M-i}z^{-1} & z^{-1} \end{bmatrix} \right\} \begin{bmatrix} 1 \\ 0 \end{bmatrix} , \tag{2.44}$$

with $r_0 = 1$. From Equation 2.44 it is clear that $A(z)$ is a polynomial in z^{-1} and according to the Equations 2.37 and 2.43 the transfer function of the synthesis filter $H(z)$ is given by

$$H(z) = \frac{1}{A(z)} . \tag{2.45}$$

A conclusion is that the synthesis filter is a recursive all-pole network and that stability of this network is guaranteed since it represents a passive acoustic tube, under the condition that $-1 \leq r_m \leq 1$, or equivalently $\mathcal{A}_m \geq 0$, for all m. This means that all poles of $H(z)$ are inside the unit circle in the z-plane. Consequently $H(z)$ represents a minimum-phase network.

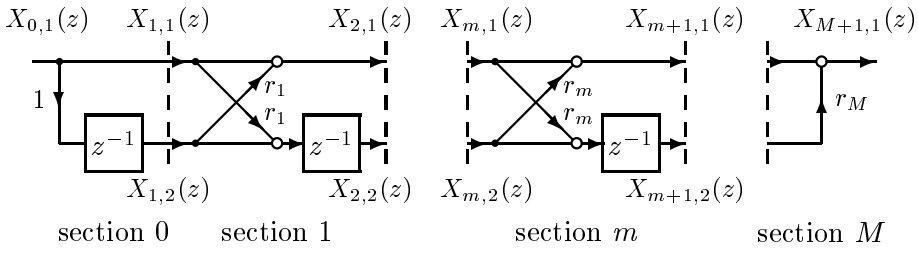

Figure 2.28: Lattice inverse filter $A(z)$.

Another result, which directly follows from the above equations, is the network structure of the inverse filter with transfer function $1/H(z) = A(z)$ with input $X_{0,1}(z)$ and output $X_{M+1,1}(z)$, as shown in Figure 2.28. Each section in this network represents one of the factors of Equation 2.44.

2.4.4 Direct-form networks

By inspection of the lattice network of Figure 2.28 with transfer function $A(z)$ of order M, which will explicitly be denoted by $A_M(z)$, we see that the impulse response $\alpha_M[n]$ of the network must be $M+1$ samples long:

$$A_M(z) = \sum_{n=0}^{M} \alpha_M[n] z^{-n} , \text{ with } a_M[0] = 1 . \tag{2.46}$$

The signal-flow graph of this direct form is shown in Figure 2.29, where

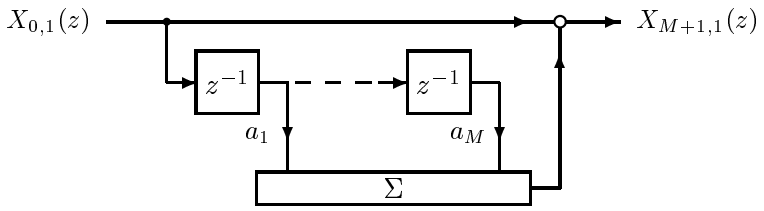

Figure 2.29: Direct-form inverse filter $A(z)$.

the coefficients are denoted by a_i. These a-coefficients, also often referred to as a-parameters, are directly related to the impulse response by

$$a_i = \alpha_M[i], \quad i = 0, 1, \ldots, M \; . \tag{2.47}$$

The transfer function

$$\frac{1}{A(z)} = \frac{1}{1 + \sum_{i=1}^{M} a_i z^{-i}} \tag{2.48}$$

is then represented by the network of Figure 2.30. Again, stability is

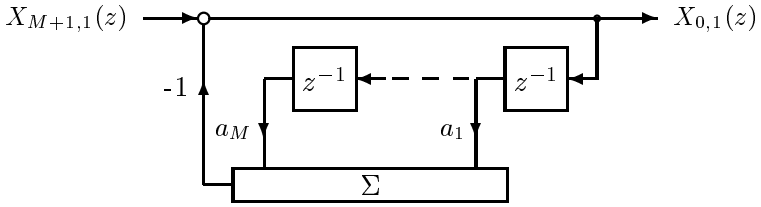

Figure 2.30: Direct-form synthesis filter $1/A(z)$.

guaranteed if the a-coefficients correspond to the reflection coefficients of an acoustic tube with positive areas.

2.4.5 Determination of the direct-form coefficients

The a-parameters can be obtained from the reflection coefficients on the basis of a recursive relation. For this purpose we define

$$A_m(z) = \frac{X_{m+1,1}(z)}{X_{0,1}(z)} \ , \quad m = 0, 1, \ldots, M \ , \tag{2.49}$$

and

$$B_m(z) = \frac{X_{m+1,2}(z)}{X_{0,1}(z)} \ , \quad m = 0, 1, \ldots, M-1 \ , \tag{2.50}$$

with reference to Figure 2.28. By inspection of section m of the same figure, we see that

$$A_m(z) = A_{m-1}(z) + r_m B_{m-1}(z) \tag{2.51}$$

and

$$B_m(z) = z^{-1} \left\{ r_m A_{m-1}(z) + B_{m-1}(z) \right\} \ . \tag{2.52}$$

By substitution of the time-flip-and-shift relation

$$B_{m-1}(z) = z^{-m} A_{m-1}(z^{-1}) \ , \tag{2.53}$$

which is proved in Appendix C, into Equation 2.51, we obtain the recursion

$$A_m(z) = A_{m-1}(z) + r_m z^{-m} A_{m-1}(z^{-1}) \ , \ m = 1, 2, \ldots, M \ , \tag{2.54}$$

where $A_0(z) = 1$ which is evident from Figure 2.28. The direct-form representation of $A_{m-1}(z)$ is given by

$$A_{m-1}(z) = \sum_{n=0}^{m-1} \alpha_{m-1}[n] z^{-n} \ , \ \alpha_{m-1}[0] = 1 \tag{2.55}$$

and its time-flipped and shifted version is

$$z^{-m} A_{m-1}(z^{-1}) = \sum_{n=1}^{m} \alpha_{m-1}[m-n] z^{-n} \ , \ \alpha_{m-1}[0] = 1 \ . \tag{2.56}$$

Hence, in the time domain the impulse response of $A_M(z)$ obeys the recursion

$$\left.\begin{array}{rcl} \alpha_m[m] & = & r_m \\ \alpha_m[n] & = & \alpha_{m-1}[n] + r_m \alpha_{m-1}[m-n] \ , \ n = 1, 2, \ldots, m-1 \end{array}\right\} \ , \tag{2.57}$$

which is iterated for $m = 1, 2, \ldots, M$.

2.4.6 The backward recursion

The backward recursion which computes the set of reflection coefficients from the a-parameters can also be formulated. We can write $A_M(z)$ as:

$$A_M(z) = \sum_{n=0}^{M-1} \alpha_M[n] z^{-n} + \alpha_M[M] z^{-M} \ , \tag{2.58}$$

and from the forward recursion, see Equation 2.57, we know that

$$r_M = \alpha_M[M] \ . \tag{2.59}$$

This shows the structure of the backward recursion. It is based on expressing $A_{M-1}(z)$ in terms of $A_M(z)$ by using r_M, and again using the last coefficient $\alpha_{M-1}[M-1]$ of $A_{M-1}(z)$ to find r_{M-1}, and so on.

By writing $B_{m-1}(z)$ explicitly from Equation 2.52 and substituting this into Equation 2.51 we find

$$(1 - r_m^2) A_{m-1}(z) = A_m(z) - r_m z B_m(z) \ . \tag{2.60}$$

By using Equation 2.53 to remove $B_m(z)$ from Equation 2.60 we find the desired relation

$$A_{m-1}(z) = \frac{A_m(z) - r_m z^{-m} A_m(z^{-1})}{1 - r_m^2} \ . \tag{2.61}$$

So, the backward recursion becomes:

$$\left.\begin{array}{rcl} r_m & = & \alpha_m[m] \\ \alpha_{m-1}[n] & = & \dfrac{\alpha_m[n] - r_m \alpha_m[m-n]}{1 - r_m^2} \ , \ n = 1, 2, \ldots, m-1 \end{array}\right\} \ , \tag{2.62}$$

which is iterated for $m = M, \ldots, 2, 1$.

2.5 Linear prediction and modelling

2.5.1 Linear prediction (LP) or LP coding (LPC)

Linear prediction (LP), or linear predictive coding (LPC), of speech is the prediction in discrete time of a current speech sample $s[n]$ on the basis of a linear combination of previous speech samples $s[n-m]$,

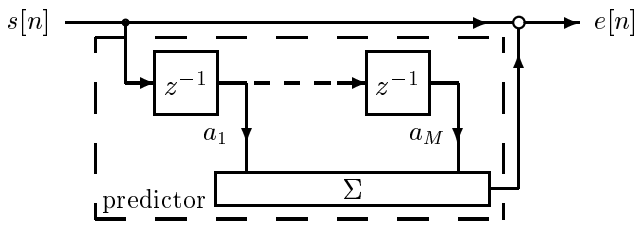

Figure 2.31: Linear prediction of the speech signal $s[n]$, 1968.

$m = 1, 2, ..., M$. The network providing the linear combination of previous samples is called the predictor, see Figure 2.31. The prediction error, or prediction residual, $e[n]$, is given by

$$e[n] = s[n] + \sum_{m=1}^{M} a_m s[n-m] \ , \tag{2.63}$$

where a_m are the prediction coefficients, or a-parameters, and M is the order of the predictor. The a-parameters are determined to minimize the total energy E of the prediction error over a certain interval $\{n_0, n_1\}$,

$$E = \sum_{n=n_0}^{n_1} e^2[n] \ . \tag{2.64}$$

To minimize E, the partial derivatives of E with respect to each a_m, $m = 1, 2, ..., M$ are set to zero. This results in M linear equations, which are usually represented in the matrix notation,

$$\mathbf{Ca} = \mathbf{c_0} \ , \tag{2.65}$$

where $\mathbf{C}$ is a symmetric $M \times M$ matrix with elements

$$c[i,j] = c[j,i] = \sum_{n=n_0}^{n_1} s[n-i]s[n-j] \ , \ i,j = 1, 2, \ldots, M \ , \tag{2.66}$$

$\mathbf{a}$ is an $M \times 1$ column vector with the unknown elements a_i, and $\mathbf{c_0}$ is an $M \times 1$ vector with elements $-c[i,0]$.

LPC and its application to speech signals is extensively discussed in the literature. The first time the term "linear prediction" arose in literature was in Norbert Wiener's work during the 1940's, first in a book

of limited circulation issued in February 1942 and later in a generally available issue from 1949 [26]. The book deals with extrapolation, interpolation and smoothing of stationary time series. In Appendix B of the book, and originally in [27], Norman Levinson described a recursive algorithm to compute the optimal a-parameters for a stationary signal, in which case it was appropriate to choose an infinitely long minimization interval $\{n_0, n_1\}$. This so-called "Levinson recursion" solved the equation $\mathbf{Ca} = \mathbf{y}$ in an efficient way where $\mathbf{C}$ is a symmetrical Toeplitz matrix and $\mathbf{y}$ is an arbitrary vector. The matrix $\mathbf{C}$ of Equation 2.65 is not Toeplitz, however. Both the Toeplitz and the non-Toeplitz cases will be addressed below.

An example of early application (on television signals) of LPC, still with fixed coefficients, was reported by Harrison already in 1952 [28]. The application of LPC to speech signals in an adaptive way, where the a-parameters are computed for subsequent segments of speech to follow the changing characteristics of the speech signal was independently introduced by Itakura and Saito [29] and by Atal and Schroeder [30] in the late sixties. Atal and Schroeder described a variant of LPC which later became known as the "covariance method". In this method the prediction residual $e[n]$ is windowed according to Equation 2.64 yielding Equation 2.65. In this case, the a-parameters can efficiently be solved by the square root or Cholesky algorithm [31] or the modified Cholesky algorithm avoiding the square roots, as described in Chapter 8 of the book of Rabiner and Schafer on digital processing of speech signals [32], but these are more complex than the Levinson recursion.

The method of Itakura and Saito was different. It was derived from a statistical point of view and they called it the PARCOR (from partial correlation) method [33]. Later, it turned out that this approach was essentially the same as what has become known as the autocorrelation method [34]. In this variant, the input speech signal $s[n]$ is windowed instead of the prediction residual - such that all data outside the window is zero - and E is determined to include all samples $e[n]$ different from zero [35,36]. This agrees with an infinitely long minimization interval again, like in Levinson's case. In the autocorrelation method, the equations take the form

$$\mathbf{Ra} = \boldsymbol{\rho} \,, \tag{2.67}$$

where $\mathbf{R}$ is a symmetric - and also Toeplitz - $M \times M$ matrix with the

elements

$$\rho[|i-j|] = \sum_{n=n_0}^{n_1-|i-j|} s[n]s[n+|i-j|] \ , \ i,j = 1,2,\ldots,M \ , \tag{2.68}$$

and $\boldsymbol{\rho}$ is a column vector with elements $-\rho[i]$, $i = 1, 2, ..., M$. Equation 2.67 can be solved with the aid of the Levinson recursion, but, because the vector $\boldsymbol{\rho}$ has the same elements as the matrix $\mathbf{R}$, the solution method can be made even more efficient. An algorithm making use of this property is known as the Levinson-Durbin recursion [32,34,36]. Just like the original Levinson recursion, it is based on expressing the solution of order m in terms of the solution of order $m-1$, and it is given by

$$E_0 = \rho[0] \tag{2.69}$$

$$r_m = -(\rho[m] + \sum_{n=1}^{m-1} \alpha_{m-1}[n]\rho[m-n])/E_{m-1} \tag{2.70}$$

$$\alpha_m[m] = r_m \tag{2.71}$$

$$\alpha_m[n] = \alpha_{m-1}[n] + r_m\alpha_{m-1}[m-n] \ , \ n = 1,2,\ldots,m-1 \tag{2.72}$$

$$E_m = E_{m-1}(1-r_m^2) \tag{2.73}$$

where Equations 2.70–2.73 are computed recursively for $m = 1, 2, ...M$. The solution is then given by $a_m = \alpha_M[m]$, $m = 1, 2, ...M$. The auxiliary variable r stands for reflection coefficient.

2.5.2 Properties of LP and relations to the acoustic tube

Indeed, the recursion of Equations 2.71-2.72 is exactly the same as the one of Equation 2.57, which was derived to compute the direct-form coefficients from the reflection coefficients of the acoustic tube. Moreover, it is observed that the linear prediction system of Figure 2.31, and the associated system $A(z) = E(z)/S(z)$ with $E(z)$ and $S(z)$ being the z-transforms of $e[n]$ and $s[n]$ respectively, corresponds to the direct-form inverse transfer function $A(z)$ of the acoustic tube according to Equation 2.46 and Figure 2.29. If the system 1/A(z) according to Equation 2.48 and Figure 2.30 is a stable system, reflection coefficients of an acoustic tube can be computed by executing Equation 2.62, and by using Equation 2.34 positive areas will be obtained. In turn, Equation 2.35 can be used to check stability of $1/A(z)$, if this is not known a priori. The link between LPC and the acoustic tube was described for the first time

by Atal and Hanauer [37]. Next, we will review some of the attractive properties of the system $A(z)$.

First of all, it is easily proved that $1/A(z)$ is always a stable system if the coefficients are computed on the basis of the autocorrelation method. In this case Equation 2.73 is satisfied which means that for any non-zero signal $s[n]$, E_m and E_{m-1} are positive non-zero quantities because they represent energy according to Equation 2.64 and consequently r_m must satisfy $r_m^2 < 1$ [34, Section 5.2.6]. This means that the all-pole transfer function $1/A(z)$ has all its poles inside the unit circle and that it is a minimum phase transfer function. Consequently, the all-zero function $A(z)$ has all its zeros inside the unit circle as well, and it is also a minimum phase network.

Stability of the covariance approach, on the other hand, cannot be guaranteed although the covariance method usually does not differ much from the autocorrelation method. In practice, the segment length $\{n_0, n_1\}$ is often 20 ms - for the same reasons as discussed in Section 2.1 - covering 160 samples at 8 kHz sampling frequency and, as we will see below, a common value of the prediction order is 10. This shows that only 10 samples out of 160 are different in both methods (compare Equations 2.66 and 2.68).

Before touching upon some spectral properties of $A(z)$, we mention another recursion to solve the Toeplitz system which is usually accredited to the work of the German mathematician J. Schur from 1917 [38], and it was reintroduced by Dewilde, Vieira and Kailath in 1978 [39]. A way of deriving the Schur algorithm is as follows. By observing the sum in Equation 2.70 we recognise a (partial) convolution of the a-parameters of a system of order $m-1$ and the sequence of correlation coefficients $\rho[i]$. This can be interpreted as filtering the sequence $\rho[i]$ by the direct-form of $A_{m-1}(z)$. Since we know from the acoustic tube analysis of Section 2.4 that $A_{m-1}(z)$ has an equivalent lattice-form according to Figure 2.28, we can rewrite the partial convolution in the lattice form. Doing so, the a-parameters in the Levinson-Durbin algorithm of Equations 2.70-2.72 become redundant parameters and they can be omitted:

$$E_0 = \rho[0] \tag{2.74}$$

$$\left.\begin{array}{rcl} r_m & = & -\beta_m / E_{m-1} \\ E_m & = & E_{m-1}(1 - r_m^2) \end{array}\right\} m = 1, 2, \ldots, M \tag{2.75}$$

In this recursion β_m equals the part between brackets of the numerator of Equation 2.70. For $m = 1$ it equals $\beta_1 = \rho[1]$. In the computation of

β_m for $2 \leq m \leq M$, we will use a denotation according to

$$\begin{aligned} p_m[n] &= x_{m,1}[n] \\ q_m[n] &= x_{m,2}[n+1] \end{aligned} \tag{2.76}$$

where $x_{m,j}$ is defined according to Figure 2.28 and $x_{m,1}[n]$ is the upper-branch output of section m and $x_{m,2}[n+1]$ is the bottom-branch output of the same section that is present *before* the delay element at time instant n. If we initialise $p_m[n]$ and $q_m[n]$ by

$$\begin{aligned} p_1[n] &= \rho[n], \quad n = 2, 3 \ldots, M \\ q_1[n] &= \rho[n], \quad n = 1, 2 \ldots, M-1 \end{aligned} \tag{2.77}$$

which provides the complete input sequence $\rho[1], \rho[2], \ldots, \rho[M]$ of the filter, we can write for section 1

$$\begin{bmatrix} 1 & r_1 \\ r_1 & 1 \end{bmatrix} \begin{bmatrix} p_1[n] \\ q_1[n-1] \end{bmatrix} = \begin{bmatrix} p_2[n] \\ q_2[n] \end{bmatrix} \tag{2.78}$$

which stands for $A_{m-1}(z)$ with $m = 2$. Thus, referring to Equation 2.70 again, we find $\beta_2 = p_2[2]$. The next step is to process the remaining input samples defined by $n = m+1, \ldots, M$, by the same section in order to prepare the required input for the next section. By continuation of this process for all sections we find the recursion

$$\left.\begin{aligned} &\begin{bmatrix} 1 & r_{m-1} \\ r_{m-1} & 1 \end{bmatrix} \begin{bmatrix} p_{m-1}[n] \\ q_{m-1}[n-1] \end{bmatrix} = \begin{bmatrix} p_m[n] \\ q_m[n] \end{bmatrix}, \; n = m, \ldots, M \\ &\beta_m = p_m[m] \end{aligned}\right\} \tag{2.79}$$

which is iterated for $m = 2, \ldots, M$. The main loops of Equations 2.75 and 2.79 can readily be combined in a straight-forward way to form one single recursion for $m = 2, \ldots, M$. This recursion must be preceded by the initiations and the computations for the case $m = 1$. Furthermore, careful analysis shows that it is redundant to use M arrays $p_m[n]$ and M arrays $q_m[n]$. If organized properly, single arrays $p[n]$ and $q[n]$ can be used. Thus the complete Schur recursion becomes

$$\begin{aligned} p[n] &= \rho[n], \quad n = 2, 3 \ldots, M \\ q[n] &= \rho[n], \quad n = 1, 2 \ldots, M-1 \end{aligned} \tag{2.80}$$

as initialisation,

$$\begin{aligned} r_1 &= -\rho[1]/\rho[0] \\ E_1 &= \rho[0](1 - r_1^2) \end{aligned} \tag{2.81}$$

for $m = 1$, and

$$\left.\begin{array}{l} \left[\begin{array}{cc} 1 & r_{m-1} \\ r_{m-1} & 1 \end{array}\right] \left[\begin{array}{c} p[n+1] \\ q[n] \end{array}\right] = \left[\begin{array}{c} p[n] \\ q[n] \end{array}\right], n = 1, \ldots, M-m+1 \\ \\ r_m = -p[1]/E_m \\ E_m = E_{m-1}(1-r_m^2) \end{array}\right\} \quad (2.82)$$

which is iterated for m=2,...,M. The Schur recursion is especially of interest when fixed-point arithmetic is used because the a-parameters are absent. The a-parameters tend to have a Gaussian amplitude distribution while the magnitudes of the reflection coefficients are conveniently bounded by unity. The algorithm was reinvented by Le Roux and Gueguen in 1977, without any reference to Schur [40]. An overview of several Levinson and Schur algorithms is found in [41].

Another property of the system $A(z)$ is that it provides maximum spectral flatness of the prediction residual $e[n]$, albeit only provable for the autocorrelation method [34, Section 6.3]. This property is also discussed in Section 3.6.1. In that section it is clearly shown that minimization of the energy of the prediction residual $e[n]$ gives rise to maximum spectral flatness of it. This means that the transfer function of $A(z)$ is approximately inverse to the spectral envelope of its input signal, if it has enough coefficients, and that the spectral shape of the input signal is represented by the prediction coefficients. The system $A(z)$ is referred to as the inverse filter, and it can be realized according to, e.g., the direct form of Figure 2.31, or the lattice form of Figure 2.28. Consequently, the function $1/A(z)$ represents the spectral envelope of the speech segment under consideration and it provides a useful synthesis structure. The function $1/A(z)$ is therefore referred to as the synthesis filter. It can be realized according to, e.g., the direct form of Figure 2.30, the lattice form of Figure 2.27, or even the ladder form of Figure 2.23.

On the basis of the foregoing paragraphs, we are able to make a good estimate of the prediction order M. Since a second-order function is required to create a formant and since there are three to four formants in narrow-band speech, six to eight coefficients are needed to realize the required formant structure. In actual operation, not all coefficients may be devoted to formants, but some of them may be needed to represent global spectral inclination. Figure 2.32 shows an example of the amplitude spectrum of a 20-ms voiced speech segment and the transfer function of an associated 10-th order synthesis filter computed

with the autocorrelation method, showing four formants.

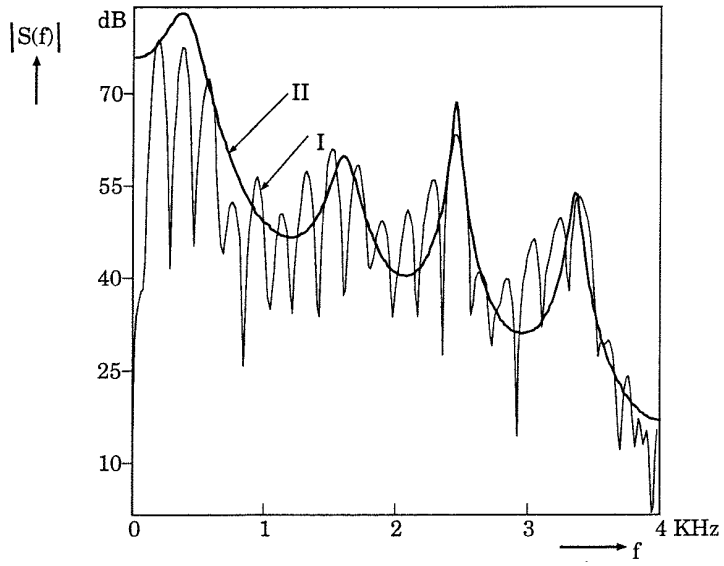

Figure 2.32: I: Amplitude spectrum of a 20-ms voiced speech segment (note the pitch harmonics). II: Absolute value of the transfer function of the associated 10-th order synthesis filter (shifted vertically to fit the envelope of the spectrum).

On the other hand one could argue that the dimensions of the associated acoustic tube of the predictor should more or less fit the physical dimensions of the vocal tract. With reference to Appendix A, where we found that the propagation speed of sound in air at body temperature equals 354 m/s and assuming a length of the vocal tract of 17 cm again, we find that the vocal tract stands for a one-way delay of about 0.5 ms. As we have seen from the acoustic tube analysis, the total delay in the predictor is twice this figure, so it amounts to 1 ms. At a sampling frequency of 8 kHz this leads to $M = 8$ which agrees very well with the first estimation. It must be taken into account, however, that the eighth-order system only suffices for modelling the vocal tract. Because the linear predictor flattens the spectrum of the *speech* signal as well as it can, the spectral properties of the excitation function and the acoustic radiation function from the mouth, causing possible spectral inclination,

are included. Some additional coefficients are needed for this purpose.

2.5.3 Modelling with LP

The excitation and radiation functions have been studied in detail as well, and there exists a rich literature on this subject. For voiced speech, the excitation waveform is characterised by a quasi-periodic pulse form which is zero for a part of the period representing glottal closure. Stevens, Kasowski and Fant (1953) described it as a sawtooth-like waveform the harmonics of which decay in amplitude with increasing frequency at 6 to 12 dB/octave [23]. An elaborate study was made by Flanagan which was described in his book, published in 1972 [8]. Especially Section 6.24 (Excitation of Electrical Synthesisers) is of interest in this respect, where a spectral decay of 12 dB/octave is indicated as the general trend of the excitation amplitude spectrum. Rosenberg conducted experiments investigating the effect of various glottal pulse shapes on the quality of speech, the results of which were published in 1972 [42]. One of the preferred shapes was defined by

$$x = \begin{cases} 3\left(\frac{t}{T_p}\right)^2 - 2\left(\frac{t}{T_p}\right)^3 & 0 \le t \le T_p \\ 1 - \left(\frac{t-T_p}{T_N}\right)^2 & T_p \le t \le T_p + T_N \\ 0 & T_p + T_N \le t \le T_0 \end{cases} , \tag{2.83}$$

with

$$\frac{T_p}{T_0} = 0.4 \text{ and } \frac{T_N}{T_0} = 0.16 , \tag{2.84}$$

where T_0 is the duration of the pitch period. The amplitude spectrum of this pulse form has also a decay of 12 dB/octave. Figure 2.33 shows this excitation waveform and its amplitude spectrum.

By and large, it can be concluded that voiced excitation can be modelled well by a second-order function which would ask for two extra coefficients in LPC on top of the formant modelling.

For unvoiced speech, often simple models are used with white noise as an excitation source for the synthesis filter. The required order of the synthesis filter for unvoiced sounds is generally lower than for voiced speech, so that an LPC system with an order tuned to voiced speech will amply cover all cases.

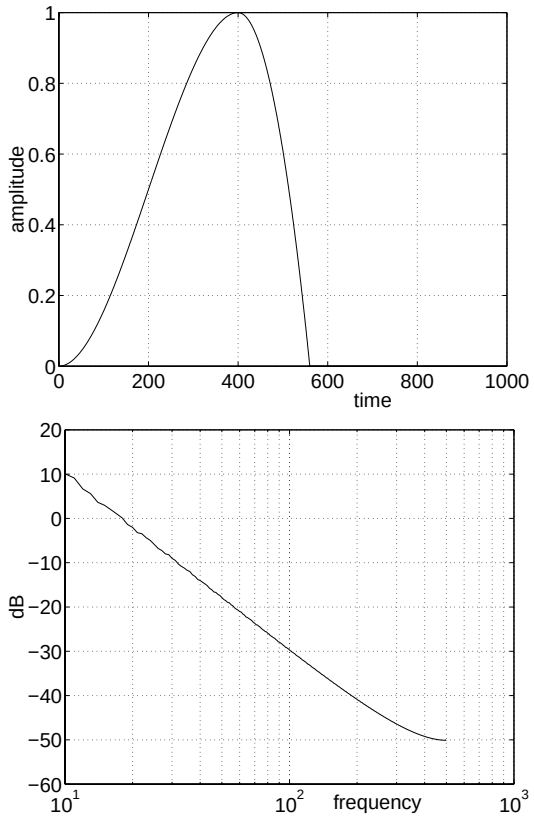

Figure 2.33: Waveform and amplitude spectrum of Rosenberg's glottal pulse model from 1972. The time axis is in number of samples. The spectrum, which is obtained by a 1000-point FFT of the waveform, is only shown over the first 500 of these points.

The radiation function is basically the relation between the volume velocity leaving the opening of the mouth and the pressure of the wave propagating into free air from that location. This is of particular relevance since the ear is basically a pressure transducer. According to Flanagan [8, Sections 3.3 and 6.25] the radiation function for most speech frequencies is adequately modelled as the radiation load on a piston in a large baffle. For frequencies less than 4 kHz, this can very well be approximated by the properties of a small, as compared to the wavelength, spherical acoustic source. This comes down to differentiation of the volume velocity wave to obtain the pressure wave [43]. Spec-

trally, this implies an amplitude increase with increasing frequency by 6 dB/octave. In Sections 2.3 and 2.4 we modelled the output of the composite acoustic tube by a large baffle, representing a last uniform-tube section with a large cross-sectional area. The electrical transmission-line model of such a section was given by the T-network of Figure 2.21. Assuming a small length of the section, it was shown to be well approximated by a T-network having two equal series inductances $Z_1 = \mathcal{L}/2$ with the midpoint connected to a shunt capacitance $Z_2 = \mathcal{C}$. Because of the large area of this last section, $\mathcal{C}$ is relatively large, representing a short circuit for most frequencies, so that the input voltage of this section - which is the output voltage of the composite tube - is mainly determined by the inductance $\mathcal{L}/2$. Fully in agreement with the above spherical acoustic source model, the output voltage of the composite tube is then obtained by differentiation of its output current.

A complete acoustic model can now be composed (see Figure 2.34). The excitation part of it consists of an impulse generator which generates a spectrally flat impulse train the pitch of which can be controlled by a pitch parameter, a glottal-pulse model in the form of the transfer function $G(z)$, and a gain factor g_v which controls the sound level, for voiced sounds. For unvoiced sounds the excitation part consists of a white noise generator, a transfer function $N(z)$ to model the spectral shape of the noise source, and a gain factor g_u. The vocal tract and the lip radiation are represented by the transfer functions $V(z)$ and $R(z)$, respectively. The system represented by $V(z)$ is excited by either the voiced or the unvoiced excitation signal, or in a more sophisticated version of the model, by a linear combination of the two excitation signals.

2.5.4 Estimation of the vocal tract parameters

When LPC is applied to the output speech signal, then the resulting inverse filter $A(z)$ will represent $A(z) = G(z)V(z)R(z)$, for voiced speech. If we deduct the +6 dB/octave spectral slope of $R(z)$ from the -12 dB/octave slope of $G(z)$, the net result of these two functions in the output speech signal will be a -6 dB/octave spectral slope. If this slope is compensated by applying LPC to the differentiated speech signal instead of the speech signal itself, then the resulting reflection coefficients will directly represent the acoustic tube $V(z)$. This conclusion, already noted by Atal and Hanauer in their 1971 paper [37], has also been used by Wakita in 1973 in his investigations of the direct estimation of the

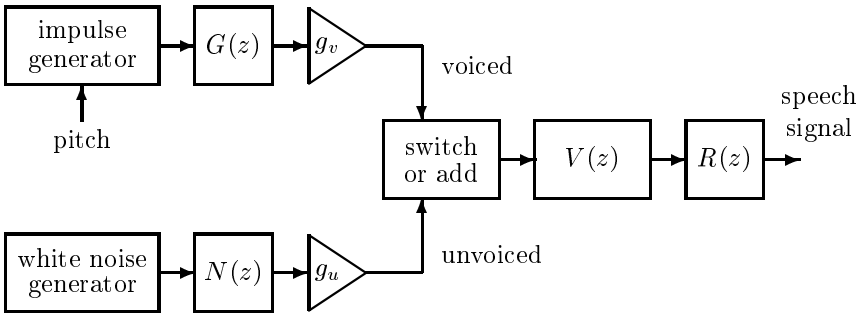

Figure 2.34: Complete acoustic model with glottal model $G(z)$, unvoiced source model $N(z)$, vocal tract model $V(z)$ and radiation model $R(z)$. The vocal tract is excited by either the quasi-periodic voiced excitation signal with gain g_v or the noisy unvoiced excitation signal with gain g_u, or by the linear combination of these signals.

vocal tract shape from the speech waveform on the basis of LPC [44]. The same path has been followed in a similar study by Markel and Gray, discussed in their book of 1976 [34]. They used a glottal model $G(z)$ represented by a second order low-pass filter with a cut-off frequency of 100 Hz and, again, a spectral slope of -12 dB/octave for higher frequencies.

Concerning the results of Wakita's work, he concluded that "...fairly reliable area functions for voiced sounds may be extracted..." and "To obtain more reliable results, it is necessary to have better ways to account for the glottal wave shape and for (acoustic energy) losses involved in speech production." Six years later, he published a paper discussing the state of the art in this area [45]. It appeared that not much progress had been made. A similar conclusion was reported by Sondhi [46].

2.5.5 Pitch prediction or long-term prediction (LTP)

Linear prediction applied in the way discussed above, also referred to as short-term prediction, effectively describes the formant structure of a speech segment, but leaves pitch-period related long-term correlation in its residual. This is shown in the example of Figure 2.35. The upper trace in this figure shows about 200 ms of a transition from voiced to

unvoiced speech (from the first part of the word "icy" spoken by a male). The second trace shows the prediction residual of a 10^{th} order inverse

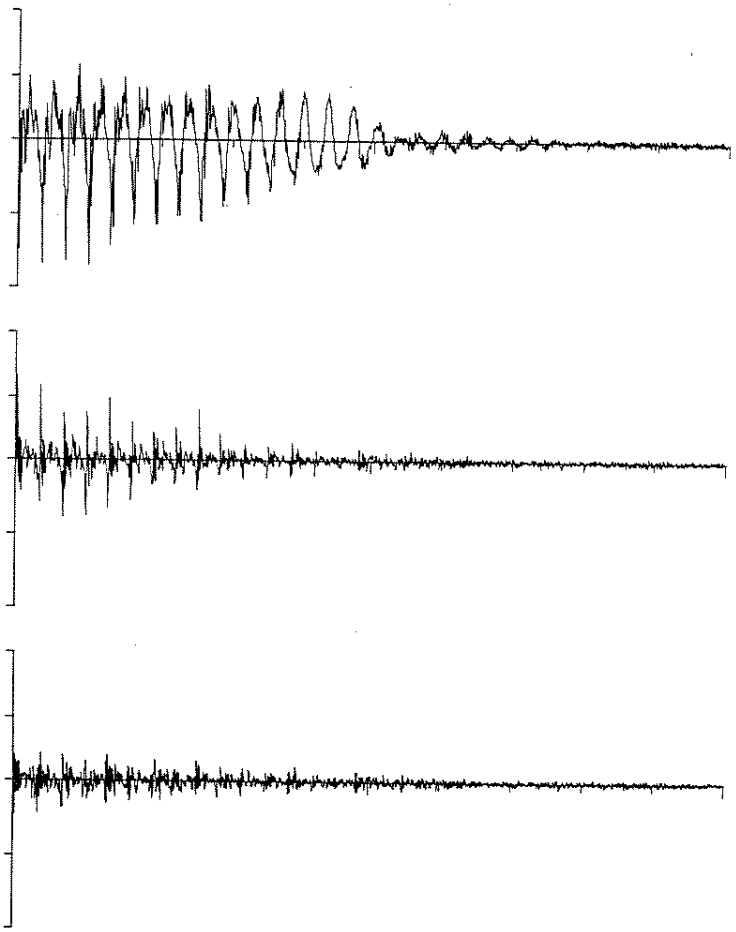

Figure 2.35: Upper trace: a portion of speech of about 200 ms; middle-trace: the short-term prediction error ($M = 10$); lower trace: the long-term prediction error. Amplitude scales are the same in all three traces.

filter, updated every 10 ms, which clearly demonstrates the presence of the long-term correlation in the form of periodic pitch pulses. These pulses could not be predicted by the short-term predictor because of their innovative nature. However, they can be predicted on the basis

of their periodicity by a pitch predictor, also called long-term predictor (LTP), according to Figure 2.36. The LTP was already proposed by

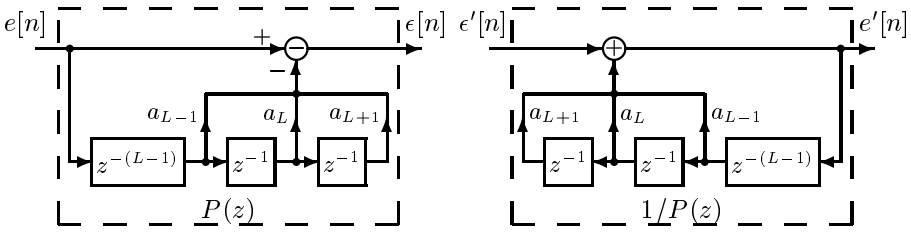

Figure 2.36: Pitch prediction analysis and synthesis filters, 1970.

Atal and Schroeder in their 1970 paper [30]. In the first-order case (only the middle coefficient) the short-term prediction residual $e[n]$ is delayed corresponding to a time lag L, weighted by a pitch-prediction coefficient a_L and subtracted from $e[n]$, resulting in the long-term prediction residual $\epsilon[n]$. The time lag, usually equal to the local pitch period, and a_L are optimised on the basis of minimising the energy of $\epsilon[n]$ and they are updated every 10 ms in the same way as in short-term prediction. It should be noted, however, that the predictor uses a subtracter here, whereas up till now we described linear predictors with an adder. The only consequence is an opposite sign for the coefficient a_L and following this convention avoids possible confusion later on. The third trace in Figure 2.35 shows the resulting $\epsilon[n]$, in which a significant reduction in dynamic range is observed.

The transfer function of such a system is referred to as $P(z)$. The network realizing the synthesis function $1/P(z)$, which restores $e'[n]$ from $\epsilon'[n]$, is also shown in Figure 2.36. If $\epsilon'[n] = \epsilon[n]$ then $e'[n] = e[n]$. The pitch synthesis filter $1/P(z)$ could be incorporated in the model of Figure 2.34, for instance by combining it with the voiced excitation generator. In that case the pitch-pulse generator should be replaced by a generator producing a more noisy excitation signal like $\epsilon[n]$.

Computation of the optimum value of the prediction coefficient can be based on the covariance method as well as on the autocorrelation method and the number of prediction coefficients can exceed one. Higher-order predictors improve the prediction, especially if the pitch period does not exactly coincide with a multiple of the sampling period. Next, we will discuss the derivation of three prediction coefficients the middle

one of which operates on the delay of L samples, the other two on $L-1$ and $L+1$. More general cases can easily be derived as well, but practical cases are mostly limited to one or three coefficients symmetrically located with respect to lag L.

The prediction error $\epsilon[n]$ is given by

$$\epsilon[n] = e[n] - \sum_{i=L-1}^{L+1} a_i e[n-i] \ , \tag{2.85}$$

and the total square error E over an N-sample segment is given by

$$E = \sum_{n=0}^{N-1} \epsilon^2[n] \ . \tag{2.86}$$

By putting the partial derivative of E with respect to a_k to zero, one finds

$$\sum_{i=L-1}^{L+1} a_i c[i,k] = c[0,k], \quad k = L-1, L, L+1 \ , \tag{2.87}$$

where

$$c[i,k] = \sum_{n=0}^{N-1} e[n-i]e[n-k] \ . \tag{2.88}$$

By back-substitution of this solution into Equations 2.85 and 2.86, one finally finds

$$E = c[0,0] - \sum_{k=L-1}^{L+1} a_k c[0,k] \ . \tag{2.89}$$

The first-order solution becomes

$$a_L = c[0,L]/c[L,L] \ , \tag{2.90}$$

with

$$E = c[0,0] - a_L c[0,L] \ . \tag{2.91}$$

The above method corresponds to the covariance method. The autocorrelation method is obtained if the input data $e[n]$ is windowed such that it becomes zero outside an N-sample segment and that the error $\epsilon[n]$ is included over all time during which non-zero samples occur. This implies that the process starts with an empty delay line, or, more precisely, with a delay line containing zero-valued samples. In this case the solution becomes

$$\sum_{i=L-1}^{L+1} a_i \rho[|i-k|] = \rho[k], \quad k = L-1, L, L+1 \ , \tag{2.92}$$

where

$$\rho[|i-k|] = \sum_{n=0}^{N-1-|i-k|} e[n]e[n+|i-k|] \; . \tag{2.93}$$

The residual error is now given by

$$E = \rho[0] - \sum_{k=L-1}^{L+1} a_k \rho[k] \; . \tag{2.94}$$

The first-order autocorrelation solution becomes

$$a_L = \rho[L]/\rho[0] \; , \tag{2.95}$$

with

$$E = \rho[0] - a_L \rho[L] \; . \tag{2.96}$$

2.6 Summary

As an introduction to this chapter, the time-domain and frequency-domain features of speech signals have been explained using the physiological model of the speech organ. The inevitable artificial source-filter model has been discussed. Next, an historical review has been presented covering ancient mechanical models, early electrical models, acoustic tube models and linear prediction.

Several ancient (18th, 19th and early 20th century) mechanical models have been described. Most of these interesting machines were based on more or less accurate duplications of the human vocal system. The attempts of von Helmholtz, from around the middle of the 19th century, to generate vowels on the basis of individual harmonic tones the amplitudes of which are adjusted to the particular vowel to be generated, were different, however. This approach can be seen as a predecessor of sinusoidal models.

Subsequently, early electrical models passed in review, starting with the model of Stewart from 1922. Then, Dudley's Voder has been described which was based on the source-filter model still in use today. Dudley's "carrier model" has been identified as another predecessor of sinusoidal models, but it got no follow-up, yet.

The contribution of Dunn has been commemorated by a review of his lossless acoustical model of the vocal tract the transfer function of which he could design by manipulation of the physical dimensions. A

detailed review of electrical analogons of these lossless acoustical models followed, building on the wave equations and their particular solutions. The understanding of these models is essential in understanding the composite acoustic-tube models and very convenient in understanding some properties of linear prediction. The section dealing with the early electrical models is concluded by a recapitulation of terminal-analog models, in which the relation to the physiological model was abandoned, once again.

Dunn's results are seen as important predecessors of the composite acoustic tube model with equal-length lossless sections. Reflection coefficients have been derived from applying the junction conditions between adjoining sections to the wave-equation solutions and it has been explained how this results in discrete-time models and lattice networks. The inverse filter has been introduced and equivalent direct-form networks have been analysed. In addition, the conversions of reflection coefficients into direct-form parameters and vice versa, have been dealt with.

The section on linear prediction (LP) or LP coding (LPC) started with the definition of the covariance and autocorrelation methods. Subsequently, the properties of LP and the direct relation to the lossless composite acoustic tube have been elaborated on, including the derivation of the Schur algorithm, analysis and synthesis filters, stability of the synthesis filter, spectral flatness of the prediction residual and the required order when applied to speech signals.

Subsequently, the use of LP in modelling has been recapitulated. A complete source-filter model with an LP synthesis filter has been described including a radiation model and an excitation function based on Rosenberg's glottal pulse model. It has been discussed how this model enables the estimation of the vocal tract dimensions directly form the speech waveform and the imperfections of this technique are pointed out. The section on LP is concluded by an outline of pitch prediction, also often referred to as long-term prediction (LTP).

The next chapter deals with the development of speech coding technology in which models of speech production play an essential role. Especially, the linear prediction technique and the related acoustic tube model form an indispensable basis for modern speech coders and in particular for the speech coder to be described in Chapter 4.

References

1. J.L. Flanagan, Voices of Men and Machines, *Journal of the Acoustical Society of America*, Volume 51, March 1972, pp. 1375–1387.
2. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978, p. 43.
3. H. Dudley and T.H. Tarnoczy, The Speaking Machine of Wolfgang von Kempelen, *Journal of the Acoustical Society of America*, Volume 22, No. 2, March, 1950, pp. 151–166.
4. W. von Kempelen, *Mechanismus der Menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*, Vienna 1791.
5. Wolfgang von Kempelen's Speaking Machine, *Philips Technical Review*, Volume 25, No.2/3, 1963/64, pp. 48–50.
6. D.J.W. Sjobbema, *Geschiedenis van de Elektronica*, Kluwer Bedrijfsinformatie B.V. Deventer, Holland, 1998.
7. R.R. Riesz, Description and Demonstration of an Artificial Larynx, *Journal of the Acoustical Society of America*, Volume 1, No. 2, January, 1930, pp. 273–279.
8. J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, Second Edition, 1972.
9. H.L.F. von Helmholtz, *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik*, sechste Ausgabe, Braunschweig Druck und Verlag von Friedr. Vieweg & Sohn, 1913, pp. 194- 201 (Künstliche Vokale). English translation: *Sensations of Tone*, Dover Publications, Inc., New York, 1954.
10. G.S. Ohm, On the definition of a Tone with the Associated Theory of the Siren and Similar Sound Producing Devices (Translated from *Poggendorf's Annalen der Physik und Chemie*, 59, 1843, pp. 497ff) in R. Bruce Lindsay (Ed.), *Acoustics: Historical and Philosophical Development*, Dowden, Hutchinson & Ross, Stroudsberg, PA, (John Wiley & Sons, Inc.), 1972, pp. 242–247.
11. J.Q. Stewart, An Electrical Analogue of the Vocal Organs, *Nature*, No.2757, Volume 110, September 2, 1922, pp. 311–312.
12. H.W. Dudley, Signal Transmission, *USA Patent 2 151 091*, March 21, 1939, Filed October 30, 1935.
13. H.W. Dudley, System for the Artificial Production of Vocal or Other Sounds, *USA Patent 2 121 142*, June 21, 1938, Filed April 7, 1937.

14. Pedro the Voder, A machine that Talks, *Bell Laboratories Record*, Volume XVII, Number VI, February, 1939, pp. 170–171. See also 'My God, It Talks', at page 177 of the same issue, and the photograph 'Pedro the Voder makes his bow at the New York World's Fair' in the May 1939 (Volume XVII, Number IX) issue.
15. The Bell System at the New York World's Fair 1939, *Bell Laboratories Record*, Volume XVII, Number X, June 1939.
16. H. Dudley, R.R. Riesz, S.S.A. Watkins, A Synthetic Speaker, *Journal of The Franklin Institute*, Volume 227, No.6, June 1939, pp. 739–764.
17. H. Dudley, The Carrier Nature of Speech, *Bell System Technical Journal*, Volume 19, 1940, pp. 495–515.
18. H. Dudley, Fundamentals of Speech Synthesis, *Journal of the Audio Engineering Society*, Volume 3, 1955, pp. 170–185.
19. H.K. Dunn, The Calculation of Vowel Resonances, and an Electrical Vocal Tract, *Journal of the Acoustical Society of America*, Volume 22, No.6, November 1950, pp. 740–753.
20. R.A.S. Paget, The Production of Artificial Vowel Sounds, *Proceedings of the Royal Society*, Volume A102, 1923, pp. 752–765.
21. I.B. Crandall, Dynamical Study of the Vowel Sounds, *Bell System Technical Journal*, Volume 6, 1927, pp. 100–116.
22. P.M. Morse, *Vibration and Sound*, McGraw-Hill Book Company, Inc., New York, 1948.
23. K.N. Stevens, S. Kasowski and C. Gunnar M. Fant, An Electrical Analog of the Vocal Tract, *Journal of the Acoustical Society of America*, Volume 25, Number 4, July 1953.
24. J.L. Kelly, Jr. and Carol C. Lochbaum, Speech Synthesis, *Proceedings of the Fourth International Congress on Acoustics*, Paper G42, 1-4 (1962); reprint in: J.L. Flanagan and L.R. Rabiner (Eds.), *Speech Synthesis*, Dowden, Hutchinson & Ross, Inc., Stroudsburg, PA (John Wiley & Sons, Inc.), 1973, pp. 127–130.
25. J.L. Flanagan, Note on the Design of "Terminal-Analog" Speech Synthesizers, *Journal of the Acoustical Society of America*, Volume 29, Number 2, February 1957, pp. 306–310.
26. N. Wiener, *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, John Wiley & Sons, Inc., New York, 1949.
27. N. Levinson, The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction, *Journal of Mathematics and Physics*, Volume XXV, No. 4, January, 1947, pp. 261–278.

28. C.W. Harrison, Experiments with Linear Prediction in Television, *Bell System Technical Journal*, July, 1952, pp. 764–783.
29. F. Itakura and S. Saito, Analysis Synthesis Telephony based on the Maximum Likelihood Method, *Proceedings of the Sixth International Congress on Acoustics*, August, 1968, Tokyo, Paper C-5-5, pp. C17–20.
30. B.S. Atal and M.R. Schroeder, Adaptive Predictive Coding of Speech Signals, *Bell System Technical Journal*, October, 1970, pp. 1973–1986.
31. D.K. Faddeev and V.N. Faddeeva, *Computational Methods of Linear Algebra*, English Translation by R.C. Williams, San Francisco: W.H. Freeman and Company, 1963, pp. 144–147.
32. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
33. F. Itakura, S. Saito, T. Koike, H. Sawabe and M. Nishikawa, An audio Response Unit Based on Partial Autocorrelation, *IEEE Transactions on Communications*, Volume COM-20, No.4, August, 1972, pp. 792–797.
34. J.D. Markel and A.H. Gray, *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.
35. J.D. Markel, Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation, *IEEE Transactions on Audio and Electroacoustics*, AU-20, June, 1972, pp. 129–137.
36. J. Makhoul, Linear Prediction: A Tutorial Review, *Proceedings of the IEEE*, Volume 63, No.4, April, 1975, pp. 561–580.
37. B.S. Atal and S.L. Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, *Journal of the Acoustical Society of America*, 50, August, 1971, pp. 637–655.
38. J. Schur, Über Potenzreihen, die im Innern des Einheitskreises beschränkt sind, *Journal für die Reine und Angewandte Mathematik*, Volume 147, No.4, 1917, pp. 205–232.
39. P. Dewilde, A.C. Vieira and T. Kailath, On a Generalized Szegö-Levinson Realization Algorithm for Optimal Linear Predictors Based on a Network Synthesis Approach, *IEEE Transactions on Circuits and Systems*, Volume CAS-25, No.9, September, 1978, pp. 663–675.
40. J. Le Roux and C. Gueguen, A Fixed Point Computation of Partial Correlation Coefficients, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, June, 1977, pp. 257–259.

41. P. Strobach, New Forms of Levinson and Schur Algorithms, *IEEE Signal Processing Magazine*, January, 1991, pp. 12–36.
42. A.E. Rosenberg, Effect of Glottal Pulse Shape on the Quality of Natural Vowels, *Journal of the Acoustical Society of America*, Volume 49, Number 2 (Part 2), 1971, pp. 583–590.
43. P. M. Morse and K.U. Ingard, *Theoretical Acoustics*, McGrawhill Book Company, Inc., New York, 1968, pp. 309–310.
44. H. Wakita, Direct Estimation of the Vocal Tract Shape by Inverse Filtering of the Acoustic Speech Waveforms, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-21, No.5, October, 1973, pp. 417–427.
45. H. Wakita, Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No.3, June, 1979, pp. 281–285.
46. M.M. Sondhi, Estimation of Vocal-Tract Areas: The Need for Acoustical Measurements, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No.3, June, 1979, pp. 268–273.

Chapter 3

The development of speech coding

In this chapter the reader is taken by the hand on a guided tour through time during which successive speech coding methods pass in review. In an original way special attention is paid to the evolutionary aspect. Specifically, for each newly proposed technique it is discussed what it added to the known techniques of the time. After presenting the relevant predecessors starting with PCM from the 1930s, we will arrive at LP-based analysis-by-synthesis systems and Regular Pulse Excitation (1984), which forms the basis of the GSM full-rate coder. Methods developed later than the mid-eighties are recapitulated in Chapter 5.

3.1 Pulse code modulation (PCM)

In attempts to transmit speech and music during the early days of telegraphy in the 19^{th} century, scientists and engineers tried to "telegraph" speech [1]. The experimenters were trying too simple codes (such as clipping the speech waveform) to obtain satisfactory quality of the reconstructed signal, however. In 1876, the invention of the telephone[1] by Alexander Graham Bell distracted the attention into the direction of analog solutions. Ironically, the title of Bell's patent reads "Improvement in Telegraphy" [2].

[1] Before Bell, the German physicist Philipp Reis succeeded in the electrical transmission of the human voice. In the period 1861–1863 he publicly demonstrated his telephone [3].

In the 1930s the Englishman Alec Harley Reeves was working in the Paris laboratories of the International Telephone and Telegraph Corporation, where he was trying to combine the advantages of telegraphy and telephony. The telegraph pulses, on the one hand, could be freed from noise to a large extent by regeneration, a process that was applied since the early days of telegraphy using electro-mechanical relays. On the other hand, telephony had the advantage of being real-time, full duplex and personal. Sampling and pulse amplitude modulation (PAM) techniques were investigated by several researchers in the early 1930s, to enable what is now known as time-division multiplex (TDM), but the pulse amplitudes remained noise sensitive. Reeves invented pulse-width modulation and pulse-time modulation and in 1937 he invented pulse-code modulation. In October 1938 he filed the first patent on PCM, consisting of the triple process of sampling, quantization and binary coding. Thus he achieved the ideal combination of telegraphy and telephony techniques. The crucial principle of regeneration could at last be applied to speech and music signals [4]. Reeves' first PCM patent was a French patent [5] which was filed in 1938, granted in 1939 and a US patent which was almost identical was granted in 1942 [6]. The first experimental equipment using PCM was built a few years later by Goodall at Bell Laboratories. It used 8000 samples per second and it was found that, with adjusted volume, a 6-bit code was necessary to provide a sufficiently low quantization noise level for good quality speech and music [7]. For obtaining a good quality with unregulated volume several additional bits were required.

In fact, the quantization noise can be expressed in terms of the amplitude probability density function (pdf) of the signal. Both for mid-rise and mid-tread uniform quantizers, the quantization error behaves as shown in Figure 3.1. If the signal x lies between the decision levels $x_i < x \leq x_{i+1}$ then the output y of the quantizer equals the reconstruction level $y = y_i$. Thus, the quantization error q is given by $q = x - y_i$.

If x is a realization of a zero-mean random variable with pdf $p_x(x)$ then its variance is given by

$$\sigma_x^2 = \mathcal{E}(x^2) = \int_{-\infty}^{\infty} x^2 p_x(x) dx \ , \tag{3.1}$$

where $\mathcal{E}$ stands for expectation. Then, q is also a random variable with

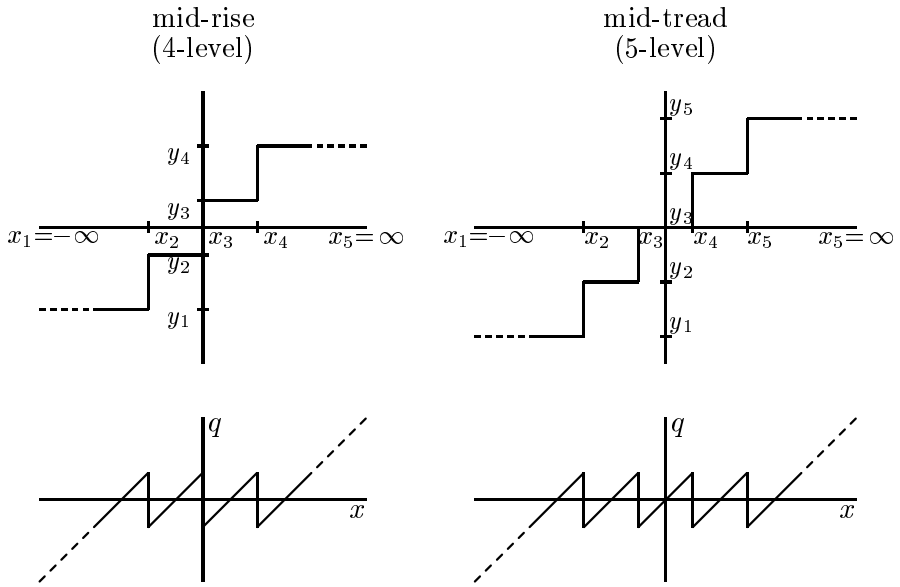

Figure 3.1: Two uniform quantizer realizations: the mid-rise and the mid-tread quantizers. The corresponding quantization errors q are shown in the bottom traces.

variance

$$\sigma_q^2 = \int_{-\infty}^{\infty} q^2 p_q(q) dq \; . \tag{3.2}$$

For a uniform quantizer with quantization step Δy and a uniform distribution of the quantization error $p_q(q) = 1/(\Delta y)$ between $q = -\Delta y/2$ and $q = \Delta y/2$, and $p_q(q) = 0$ elsewhere, it follows straightforwardly from Equation 3.2 that

$$\sigma_q^2 = \frac{(\Delta y)^2}{12} \; . \tag{3.3}$$

So, if one bit is added the quantization step Δy can be halved, giving the well known rule of thumb of 6 dB/bit SNR for a uniform quantizer.

From the struggle against transmission noise in analog telephony the merits of amplitude compression at the transmitting side and subsequent expansion at the receiving side ("compansion") were already known since the invention of the "compandor" by Mathes and Wright in 1934 [8]. So,

very soon after the invention of PCM companding was used to reduce the effect of quantization noise. In the conventional analog compandors instantaneous companding required at least twice the original transmission bandwidth to enable faithful expansion in the receiver and in addition, distortions in the transmission path could hardly be tolerated for the same reason. Therefore, they were controlled at a syllabic rate, in which case these problems did not arise. In PCM, instantaneous companding could be used without penalty because the transmission bandwidth is determined by the bit rate and not by the analog signal contents and transmission distortion is inherently not an issue. The combination of PCM with instantaneous companding results in non-uniform quantization characteristics. It was applied for the first time in the early 1940s in a speech coder called "Sigsaly" which will be discussed in Section 3.3.1, but the first publication on the subject was by Meacham and Peterson of Bell Laboratories in early 1948 [9]. In the experimental speech transmission system they built, they could use a non-uniform (they called it "tapered") quantizer with only 7 bits resulting in a bit rate of 56 kbit/s to attain toll quality speech and music without any volume regulation. A sampling rate of 8 kHz had already become a common choice. By the end of 1948 Oliver, Pierce and Shannon gave PCM its theoretical basis [10]. The potential of PCM was proven, once and for all.

Non-uniform quantization was generalized by Lloyd and Max. Lloyd prepared a paper on the subject in 1957 but never published it until 1982 [11]. Max formulated similar analysis and published his work in 1960 [12]. The quintessence of this work lies in formulating the problem in terms of a least squares problem. If the signal x lies between the decision levels $x_i < x \leq x_{i+1}$ again and the output of the quantizer equals the reconstruction level y_i again then the quantization error $q = x - y_i$ is a function of x only, in this interval. So, we can write that the contribution d_i of this interval to the total distortion is given by

$$d_i = \int_{x_i}^{x_{i+1}} (x - y_i)^2 p_x(x) dx \ , \tag{3.4}$$

and that the total distortion σ_q^2 is consequently given by

$$\sigma_q^2 = \sum_{i=1}^{N} \int_{x_i}^{x_{i+1}} (x - y_i)^2 p_x(x) dx \ , \tag{3.5}$$

where N stands for the number of quantization intervals (requiring

$\log_2 N$ bits of code). By putting the partial derivative of σ_q^2 with respect to a particular x_k to zero, we find the optimum value

$$x_{k,opt} = (y_k + y_{k-1})/2 \ , \tag{3.6}$$

and by putting the partial derivative of σ_q^2 with respect to y_k to zero we find

$$y_{k,opt} = \frac{\int_{x_k}^{x_{k+1}} x p_x(x) dx}{\int_{x_k}^{x_{k+1}} p_x(x) dx} \ . \tag{3.7}$$

These equations tell us to choose the decision levels halfway the reconstruction levels and a reconstruction level as the centroid of the area of the pdf under $p_x(x)$ between the two decision levels. These requirements have no analytical solutions and they must be determined numerically in an iterative way. In the literature quantization tables can be found for the usual Gaussian, Laplace and Gamma pdf's [15].

Later, in 1972, PCM with a logarithmic non-uniform quantizer was standardized by the ITU and it was, and still is, widely applied [13]. Theoretical and practical backgrounds of logarithmic quantization are discussed thoroughly in [14] and [15]. In Chapter 4 of the latter reference it is argued that the performance of a logarithmic quantizer is not optimised to a particular, signal-specific pdf, but that it is very robust for a range of different pdf's. The basic idea behind logarithmic quantization is that if the input x, and its compressed version c are related by $c = \ln(1 + x)$, then the difference quotient is given by

$$\Delta c = \Delta x/(1 + x) \ . \tag{3.8}$$

If c is uniformly quantized with a quantization step $\Delta c =$ constant, and if in addition $x >> 1$ it follows from Equation 3.8 that also $\Delta x/x \simeq$ constant. This means an approximately constant relative quantization error and, consequently, a constant SNR independent of the amplitude of x. In Figure 3.2 the SNR of an 8-bit logarithmic quantizer versus the signal amplitude is sketched, for a sinusoid, together with the SNR performance of an 8-bit uniform quantizer. The ITU standard has a European A-law version and a μ-law version which is used in North America and Japan. The actual quantization tables are piecewise-linear

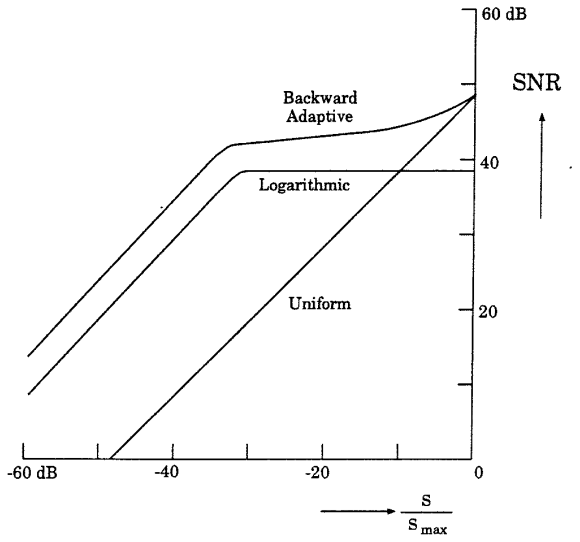

Figure 3.2: Stylised SNR performance of three different 8-bit quantizers as a function of the relative signal amplitude s/s_{max}, for a sinusoid (after[33]).

realizations of

$$c_A(x) = \begin{cases} \operatorname{sgn}(x)\frac{Ax_n}{1+\ln A}x_{\max} & 0 \le x_n \le \frac{1}{A} \\ \operatorname{sgn}(x)\frac{1+\ln(Ax_n)}{1+\ln A}x_{\max} & \frac{1}{A} < x_n \le 1 \end{cases} \tag{3.9}$$

for the A-law with A=87.56, and

$$c_\mu(x) = \operatorname{sgn}(x)\frac{\ln(1+\mu x_n)}{\ln(1+\mu)}x_{\max} \tag{3.10}$$

for the μ-law with $\mu = 255$, where x_n stands for the normalised input signal

$$x_n = |x|/x_{\max}\ , \tag{3.11}$$

and $x_{\max}$ for the amplitude of the input signal causing full load of the quantization characteristic without overload ($|q| \le \Delta c/2$).

Non-instantaneous compansion, such as applied in the conventional syllabic rate compandor, in combination with quantization is usually

referred to as "adaptive quantization". There exist two different approaches to adaptive quantization, i.e. forward adaptive quantization and backward adaptive quantization.

In forward adaptive quantization, see Figure 3.3 *a*, the level of the signal x is estimated as a function of time and a control signal representing the envelope of the signal is generated which is intended to keep the variable quantizer fully loaded while avoiding overload. As we have already seen in Figure 3.2, maximum SNR performance is obtained with fully loaded quantizers. The control signal must be transmitted together with the compressed and quantized signal. At the receiving side the

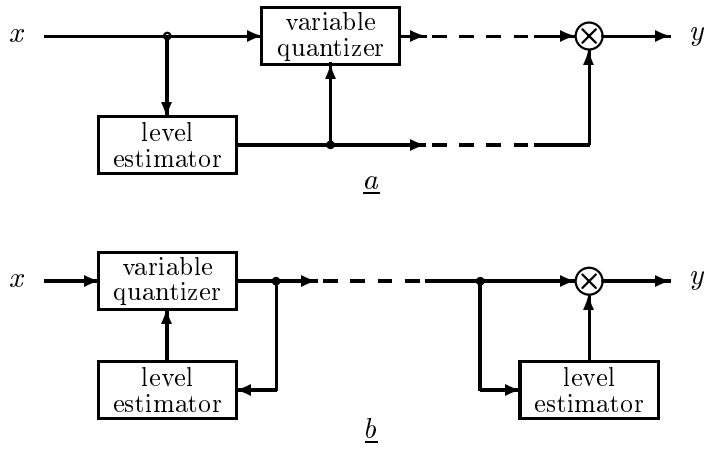

Figure 3.3: Adaptive quantization: *a*) forward adaptation and *b*) backward adaptation.

signal is expanded again by multiplying it by the envelope. So, the effective compression function of the adaptive quantizer resembles division of the signal x by its envelope.

In backward adaptive quantization, as shown in Figure 3.3 *b*, the signal level is estimated from the quantized signal. In this way it is not necessary to transmit the level information to the receiving side because it can be generated locally from the received signal. Figure 3.2 shows the performance of an 8-bit backward-adaptive (uniform) quantizer relative to the (fixed) uniform and logarithmic quantizers, for a sinusoidal signal. The level estimator approximates the signal level by observing previous quantizations. If the quantized signal samples have relatively

high amplitudes, with respect to the quantizer range, the step size in the quantization of the next signal sample will be increased by some predetermined factor and if the previous samples have a low amplitude the next step size will be decreased by some other predetermined factor. The actual factor is chosen from a fixed set, determined by the signal characteristics, dependent of the previous quantizations. A successful quantizer, developed by Jayant, works with only one-sample memory [16].

A parameter to be chosen in both forward and backward adaptation is the speed of adaptation, which should be tuned to the rate of change of the envelope of the signal. For speech, the corresponding "syllabic" time constant can vary from a few to several tens of milliseconds, preferably with a faster "attack" time and a slower "decay" time. In so-called forward block adaptation, an implementation which is especially appropriate in a digital environment, the level estimator operates on blocks of consecutive signal samples with similar durations.

Adaptive quantization was first applied by researchers in the area of differential coding, which is the subject of the next section. The application to PCM, sometimes referred to as adaptive PCM (APCM), was first reported by Golding in 1967. It concerned a study of backward adaptive quantization on television and spacecraft engineering sensor signals [17]. The first application of backward adaptive PCM to speech signals was reported by Wilkinson in 1971 [18]. The application of forward adaptation to PCM appeared for the first time in the literature in 1973 for music signals [19] and in 1974 for speech signals [20]. Further elaboration on adaptive quantization can be found in [15] Section 4.10: Adaptive Quantization. Examples of present applications are in the ITU standard G.726 (ADPCM) - see Section 1.4 and the next section - where a 4-bit backward adaptive quantizer is used and in the GSM full-rate coder - see Chapter 4 - which incorporates forward block adaptation on blocks of 5 ms in the 3-bit quantization of the excitation signal.

3.2 Differential Coders

3.2.1 Delta modulation

In delta modulation the difference between a sample and its predicted value is quantized using only one bit. In order to obtain comparable performance to PCM, the sampling frequency in delta modulation is N times higher than in N-bit PCM, resulting in approximately the same

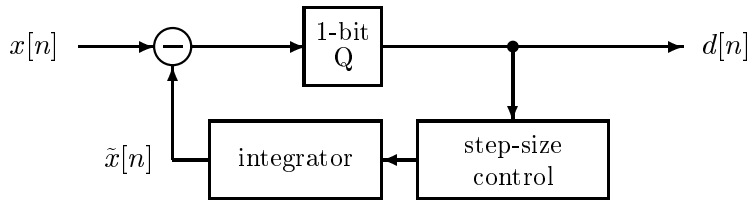

Figure 3.4: Circuit diagram of an adaptive delta modulation system with integrator, a one-bit quantizer Q and quantization step size control.

bit rate. Figure 3.4 shows an adaptive delta modulation (ADM) system. The predictor consists of an integrator and its output $\tilde{x}[n]$ pursues the input signal $x[n]$. If the magnitude of $\tilde{x}[n]$ is smaller than that of $x[n]$ then the quantizer Q decides to generate a pulse with a sign such that the magnitude of the output of the integrator will increase. If it is greater, a pulse with the opposite sign is generated to decrease it. The step-size control is a form of adaptive quantization and it determines the pulse amplitudes so as to make the pursuit as accurate as possible. The decoder receives the sequence $d[n]$ and it consists of the same step-size control and integrator to reproduce $\tilde{x}[n]$. Finally, this signal is refined by filtering it according to its original bandwidth.

Figure 3.5 shows an example of a waveform $x[n]$ and its predicted waveform $\tilde{x}[n]$. For clarity, the discrete-time samples of the waveforms have been interconnected in the plots and it is assumed that the delta modulator is provided with an ideal integrator. Practical systems always have a "leaky" integrator to prevent unlimited transmission-error propagation in the decoder, amongst other reasons. The step size control shown is due to Winkler [29], in which each successive step with the same sign is increased by a factor of two, except for two steps with equal signs following a change in sign, which have the same magnitude. If the steps alternate in sign, the magnitude of each step is half the previous one. Of course, the step-size is bounded by an absolute maximum and minimum. During the rising of the example waveform $x[n]$, the signal $\tilde{x}[n]$ cannot keep up with the slope of $x[n]$. This phenomenon is known as slope overload. The oscillation around the input waveform is called

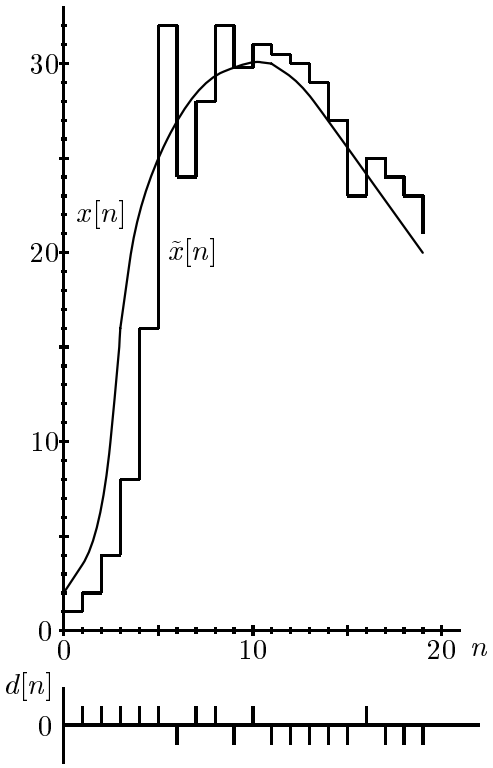

Figure 3.5: Example of a waveform $x[n]$ and the predicted waveform $\tilde{x}[]n$ (after [33]), using an ideal integrator and Winkler's step-size control [29]. For clarity, the discrete-time samples of the waveforms have been interconnected.

granular noise. In speech, slope overload is perceived as distortion and the granular noise is perceived as white Gaussian noise.

Delta modulation in its most elementary form, that is with a single integrator and without step-size control, was invented by Deloraine, Van Mierlo and Derjavitch of International Standard Electric Corporation in 1946 [21,22] and a more elaborated patent was filed in 1948 by Schouten, de Jager and Greefkes of Philips Gloeilampenfabrieken [23,24]. The first journal article on elementary delta modulation was published by the latter authors in 1952 [25]. The first improvement on the elementary delta modulation system was made by de Jager, who proposed double integration by which $\tilde{x}[n]$ became a more accurate approximation of the signal $x[n]$, and a method of modification of this double integrator to avoid

stability problems in the feedback loop [26].

A delta modulator cannot be loaded equally for all frequencies. This is most clear for an elementary delta modulator, in which the maximum slope of $\tilde{x}[n]$ occurs when all pulses $d[n]$ have the same sign. For a full-load (just no slope overload) sinusoidal input, for instance, this means that at increasing frequency the amplitude has to be decreased accordingly to avoid slope overload. This fits the decaying speech spectrum very well. A variant that can be loaded equally for all frequencies is called sigma-delta modulation. It was proposed by Inose, Yasuda and Murakami of the University of Tokyo in 1962 [27] for signals with a flat power spectrum. In sigma-delta modulation another integrator is placed at the input of the delta modulator. By replacing the two equal integrators at the inputs of the adder by a single one at the output of the adder, a sigma-delta modulator is obtained. The decoder only contains a filter corresponding to the original signal bandwidth with a flat passband. Later, sigma-delta modulation became a key-component in DA conversion of digital audio [28].

Step-size control in delta modulation was proposed by Greefkes in 1961 [30,31]. It concerned a form of forward adaptation which was called "continuous" delta modulation. At the input of the delta modulator a level estimator generated the envelope of the speech waveform in a band of 0-200 Hz. The speech signal itself was high-pass filtered above this frequency, such that the two signals were separated in frequency bands. The sum of these signals was fed to the delta modulator. By applying a low-pass filter on the binary sequence $d[n]$, both in the encoder and the decoder, the envelope was regained and applied for step-size control.

The first backward adaptive step-size controller in delta modulation was the method described above, introduced by Winkler of the Radio Corporation of America (RCA) in 1963 under the name High Information Delta Modulation [29].

An overview of the state of the art in delta modulation in the late 1960s is found in an article by Schindler of IBM [32]. In this article, four delta modulators of Philips, Bell Telephone Laboratories (both with forward adaptation), Nippon Electric and IBM (both with backward adaptation) are discussed and compared. A disadvantage of Winkler's and Schindler's methods was their nearly instantaneous character. This caused considerable granular noise and serious error propagation in the decoding process. In 1970, Greefkes and Riemens of Philips Gloeilampenfabrieken published a backward adaptation method with a syllabic

time-constant under the name Digitally Controlled Delta Modulation (DCDM) [33].

For a while, delta modulation has been considered by some researchers as an alternative candidate for the upcoming ITU standard G.711 to be fixed in 1972, but it has never been a serious threat for the logPCM systems. Later, delta modulation with another kind of step-size control, called Continuous Variable Slope Delta modulation (CVSD), has been applied in the space shuttle program [34], and in military telecommunication networks [35,36]. A late innovation, which marks the end of an era in this field, was reported by Un and Lee in 1980, who proposed the use of both syllabic and instantaneous companding [37].

3.2.2 Differential PCM

Differential PCM (DPCM) has been invented by Cutler of Bell Telephone Laboratories in 1952 [38]. It concerned a speech coder according to Figure 3.6, but with a very simple feed-back circuit consisting of an integrator, just as in delta modulation. The difference with delta mod-

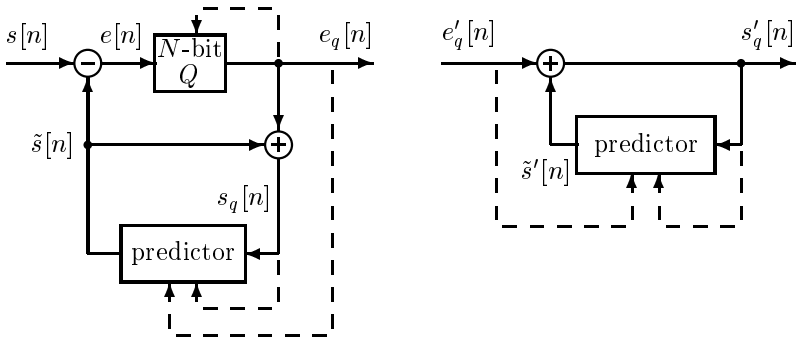

Figure 3.6: DPCM encoder (left) and decoder (right). Dashed lines indicate adaptation.

ulation is that there is no oversampling and that the N-bit quantizer Q can have more than two reconstruction levels. The filing date of Cutlers patent is June 1950. The only published material on delta modulation by that time was the very first patent of Deloraine, Van Mierlo and Derjavitch, in French (see Section 3.2.1). So, it is not unthinkable that

Cutler did not know about delta modulation.

The adder functions of Figure 3.6 are defined by

$$e[n] = s[n] - \tilde{s}[n] \ , \tag{3.12}$$

and

$$s_q[n] = e_q[n] + \tilde{s}[n] \ . \tag{3.13}$$

If the quantizer would be replaced by a short circuit then $e_q[n] = e[n]$ and substituting this into the above equations yields $s_q[n] = s[n]$. Thus, the signal $s_q[n]$ can be considered as the quantized version of the input speech signal $s[n]$. Furthermore, $\tilde{s}[n]$ is the predicted value of $s[n]$ from previous values of $s_q[n]$ and consequently $e[n]$ is the prediction error or prediction residual. In the decoder, the corresponding signals are indicated by $s'_q[n]$, $\tilde{s}'[n]$ and $e'_q[n]$, which are identical to those without prime in the encoder if $e'_q[n] = e_q[n]$. By removing $\tilde{s}[n]$ from Equations 3.12 and 3.13 we can write

$$s_q[n] = s[n] - q[n] \ , \tag{3.14}$$

where $q[n]$ stands for the quantization error

$$q[n] = e[n] - e_q[n] \ . \tag{3.15}$$

From Equations 3.14 and 3.15 we see that the quantization error in the signal $s_q[n]$ is given by the quantization error $q[n]$ in the prediction residual. This leads to the conclusion that the better the predictor works, the smaller the prediction residual will be and consequently, the smaller the range of the quantizer can be. So, improving the predictor, reducing the quantizer range and keeping the number of quantization levels constant, gives rise to smaller quantization steps and consequently, to an improved SNR. This property is quantified by the prediction gain and it is the major attribute of DPCM. Following the line of reasoning of Atal and Schroeder in their 1970 paper [39], the SNR of the reconstructed signal is given by

$$SNR = \frac{\mathcal{E}(s^2[n])}{\mathcal{E}(q^2[n])} = \frac{\sigma_s^2}{\sigma_q^2} \ , \tag{3.16}$$

where $\mathcal{E}$ stands for expectation and σ^2 for variance. Equation 3.16 can be rewritten as

$$SNR = \frac{\sigma_s^2}{\sigma_e^2}\frac{\sigma_e^2}{\sigma_q^2} = G_P.SNR_Q \tag{3.17}$$

M	a_1	a_2	a_3
1	0.8456		
2	1.3435	-0.5888	
3	1.4399	-0.8089	0.1638

Table 3.1: Prediction coefficients based on long-term speech statistics [40].

where G_P stands for the prediction gain and SNR_Q for the SNR of the quantizer itself.

After using a simple integrator for prediction in the early days of Cutlers invention, the predictor was improved by applying M-th order linear predictors of the usual form

$$\tilde{s}[n] = \sum_{i=1}^{M} a_i s_q[n-i] \ , \tag{3.18}$$

($M = 1$ is the integrator case) and by determining the predictor coefficients using a minimum mean-square prediction error criterion based on long-term speech statistics. Paez and Glisson presented such predictors in a publication in 1972 [40], see Table 3.1. The average prediction gain that can be obtained by at least a second-order predictor is about 7 dB [15], although it can be much higher in voiced sounds with prevailing low-frequency content and much lower in less predictable signals such as in most unvoiced sounds. Higher prediction orders than three hardly pay off.

Adaptive prediction was introduced in DPCM by Atal and Schroeder in 1970 [39]. It concerned forward adaptation of the predictor, that is computation of the prediction coefficients from input speech segments, so that the prediction gain is always optimised on the basis of the actual speech properties. This approach requires transmission of the coefficients to the decoder to enable appropriate decoding. Atal and Schroeder even included a pitch predictor in their work, which was the first time that pitch prediction appeared in the literature. As reported later by Jayant and Noll, see Section 6.5 of [15], adaptation can improve the average prediction gain by another 6 dB (without pitch prediction) if the order is high enough ($M \geq 10$). It should be noted that the predictor in Figure 3.6 is defined in a different way as in Section 2.5. In contrast

to Equation 2.63 the prediction error in the DPCM coder is given by

$$e[n] = s[n] - \sum_{i=1}^{M} a_i s_q[n-i] \ . \tag{3.19}$$

Apart from the difference that the prediction is based on quantized speech samples instead of the original speech samples, the a-parameters also have opposite signs. This is a result of following the convention in delta modulation and DPCM that the prediction error is defined as the *difference* between the input and the predicted speech samples. In actual operation, the opposite signs of the a-parameters compensate for the subtraction in determining the prediction residual so that the associated inverse filters $A(z)$ will be the same in both cases, apart from the quantization effects, of course.

In backward adaptation of the predictor, the a-parameters do not have to be transmitted because they can be estimated from the quantized speech signal s_q and the quantized prediction residual $e_q[n]$, both in the encoder and the decoder. In Figure 3.6 this is indicated by dashed lines. This possibility has been introduced by Gibson, Jones and Melsa in 1974 [41]. It is based on the fact that the quantized prediction residual can be minimized by adaptively driving the partial derivatives of the total square error with respect to the a-parameters, to zero. If the error measure E is taken over the interval n_0, n_1 it is defined by

$$E = \sum_{n=n_0}^{n_1} e_q^2[n] = \sum_{n=n_0}^{n_1} \left(s[n] - \sum_{i=1}^{M} a_i s_q[n-i] - q[n] \right)^2 \ , \tag{3.20}$$

and the partial derivatives are given by

$$\frac{\delta E}{\delta a_j} = -2 \sum_{n=n_0}^{n_1} e_q[n] s_q[n-j] \ , \ j = 1, 2, \ldots, M \ , \tag{3.21}$$

if we assume that $q[n]$ is independent of a_j for all $j = 1, 2, \ldots, M$ for a fixed quantizer. So, driving the cross-correlation function of Equation 3.21 into the direction of zero improves the prediction. The prediction becomes optimum if all M correlations equal zero.

In addition to adaptive prediction, the quantizer can be optimised by making it adaptive (in Figure 3.6 indicated by a dashed line). A backward adaptive quantizer was already proposed in 1955 by Cutler [42], the original inventor of DPCM. In 1973, Cummiskey, Jayant and Flanagan proposed a refined backward adaptive quantizer [43, 44].

In 1984 the ITU-T standardized Adaptive DPCM (ADPCM), both with backward adaptive quantization and backward adaptive prediction [45]. This standard was dubbed G.721 but in 1990 it has been renamed into G.726 (see Section 1.4). The main application mentioned there is DCME as described in Section 1.3.1, but it is also standardized by the ETSI for DECT cordless telephony and it has found a widespread application in voice mail systems. ADPCM is also applied in the wide-band ITU standard G.722 from 1988, operating at 48, 56 or 64 kbit/s.

If we assume the quantization noise $q[n]$ generated by the quantizer to be spectrally flat (the PCM case), then it follows immediately from Equation 3.14 that the quantization noise in the reconstructed signal $s'_q[n]$ is also flat in the DPCM case (provided that $s'_q[n] = s_q[n]$, of course). Other forms of DPCM are open-loop DPCM, also referred to as D*PCM, and Noise Feedback Coding (NFC). These variants have different properties characterised by the spectral distribution of the quantization noise. D*PCM consists of quantization of the prediction error $e[n]$ obtained by applying the inverse filter $A(z)$ to the speech signal as in Figure 2.31. The appropriate decoder then consists of the synthesis filter $1/A(z)$. Although there are early predecessors of D*PCM, see for instance [46], the combination with adaptive prediction was considered for the first time by Dunn in 1971 [47]. If we assume the quantization noise generated by the quantizer to be spectrally flat again, then the output quantization noise of the decoder will have the same spectral shape as the (adaptive) transfer function of $1/A(z)$ and, consequently, also as the speech signal itself. In a perceptual sense, this proportional noise shaping is an advantage since this can give rise to better masking of the quantization noise by the decoded speech signal [49]. Objective evaluation was presented in a paper by Noll [50] in 1978, in which he compared PCM and other forms of DPCM on the basis of SNR. He showed that the SNR of D*PCM is always less than that of DPCM. In fact, using the minimum mean square error predictor according to Section 2.5, resulting in the optimal inverse filter $A(z)$ with maximum spectral flattening, gives an SNR for D*PCM identical to that of PCM. D*PCM performs best using a half-whitening inverse filter, fully in accordance with the matched-filter technique from communications theory (see for instance [51]), resulting in a gain over PCM. Half-whitening means that the inverse filter does not have a transfer function according to the inverse (envelope) of the speech spectrum, but according to the square root of it.

NFC consists of quantization of the prediction error $e[n]$ according to Figure 3.7. By feedback of the quantization noise via a transfer function

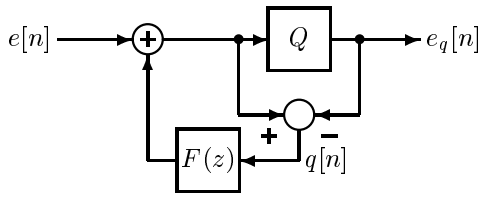

Figure 3.7: Circuit diagram of a noise feedback coder with quantizer Q and noise shaper $F(z)$.

$F(z)$, extra freedom is obtained in spectral shaping of the decoder output quantization noise. The appropriate decoder consists of the synthesis filter $1/A(z)$ again. NFC has also early predecessors [52, 53] but the combination with adaptive prediction has been considered for the first time by Noll in his 1978 paper. Straightforward analysis of this system shows that the transfer function $H_n(z)$ for the quantization noise to the output of the decoder is given by

$$H_n(z) = (1 - F(z))/A(z) \; . \tag{3.22}$$

The case $F(z) = 0$ yields a D*PCM system and the case $F(z) = 1 - A(z)$ turns into an equivalent of the DPCM system. In general, NFC provides the freedom to optimize the quantization noise in a perceptual sense. The perceptual optimum has a spectral shape somewhere between flat (as in DPCM) and proportional to the input speech signal (as in D*PCM with maximum spectral flattening), according to Atal and Schroeder in their 1979-paper [49].

3.3 Vocoders

3.3.1 Channel Vocoders

The very first speech coder, in the sense of a transmitter and a receiver facilitating efficient transmission, was the *parametric* channel vocoder of

Homer William Dudley of Bell Telephone Laboratories (see also Subsection 2.3.2). The first patent in which he described this technique was filed in 1935 and it was issued in 1939 [54]. The first publication on the subject was in 1936 [55] and the word "vocoder" was used for the first time in a paper in 1939 [56]. Remarkable about the channel vocoder is that there were no predecessors, it was fundamentally new and it had absolutely unique attributes.

The transmitter of the system described in the above documents contains an analyser that uses a pitch and voicing detector and a filterbank creating ten spectral channels covering the frequency range of from 250 to 7100 Hz according to the patent and from 0 to 2950 Hz according to the 1939-paper. In each channel the envelope of the signal as a function of time is detected and these parameters are each represented in a band of 0 - 25 Hz. In Figure 3.8 the diagram of such a system

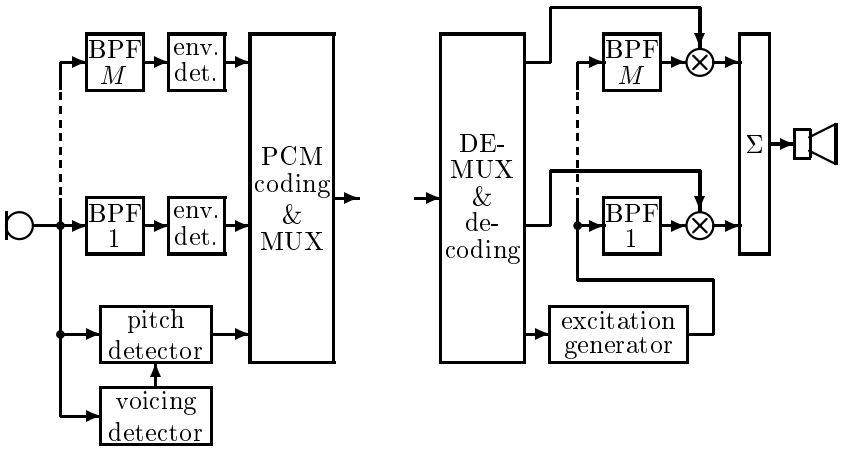

Figure 3.8: Block diagram of a channel vocoder.

is shown, including digitisation of the parameter signals. In Dudley's patent, however, the analog parameter signals are as yet combined in an analog FDM way, thus compressing the original bandwidth into a frequency band of 10 - 360 Hz. The pitch detector makes use of a frequency meter in the low-frequency region (typically 0 - 300 Hz) of the speech signal. It renders a control signal following the time-varying pitch parameter in a 0 - 25 Hz frequency band, as well. Unvoiced sounds have

not enough power in the low-frequency region to activate the frequency meter, according to Dudley, so that "zero pitch frequency" indicates these sounds. An explicit voicing detector, as shown in Figure 3.8, was not present in Dudley's first vocoder.

The synthesiser in the receiver contains an excitation generator which generates pitch pulses the frequency of which is controlled by the received pitch parameter, in the voiced case. In the unvoiced case a white noise signal is generated. Either of the two excitation signals is fed to the common input of another filter-bank, which is identical to the one in the analyser, and the appropriate time-envelope of the output of each band-pass filter (BPF) is reconstructed by means of the received spectral parameters.

In this way voiced speech with the right pitch and unvoiced speech are reproduced, both with adequate spectral features as a function of time. The original speech waveform, however, is lost, and cannot be reproduced.

The first channel vocoder with digitised parameters and digital transmission concerned a military (USA Army) secure voice communication system called "Sigsaly" built in the early 1940s [57]. It had an additional innovative digital enciphering capability, also invented at Bell Telephone Laboratories, which made it the first reliable secure voice communication system ever built. The system has been used between the Pentagon in Washington and London during the second world war from 1943 till 1946. Terminals were also installed at other important centers including Paris, North Africa, Australia, Hawaii, and the Philippines. In the mentioned period, over 3000 secret conferences have been supported by Sigsaly [58]. The existence of the system has been kept secret until 1976.

Sigsaly was really bulky, see Figure 3.9. It occupied over 30 man-high rack mounting bays. It required 30 kW of power to encode the speech signal at a bit rate of about 1.6 kbit/s, to transmit it over transatlantic short-wave radio, and to receive and decode the incoming signal to produce 1 mW of poor quality speech.

The low bit rate was achieved by using a 36-level (uniform) quantizer for the pitch parameter, which was sampled at 50 Hz, and a 6-level non-uniform quantizer with 6 dB per step for each of the ten spectral parameters, which were also sampled at 50 Hz. The logarithmic quantizer had been invented during the design period of Sigsaly by Eugene Peterson, according to William Bennett [59].

After the war, the concept of the channel vocoder was further elabo-

Figure 3.9: Sigsaly. (From: J.V. Boone and R.R. Peterson, *The Start of the Digital Revolution: SIGSALY Secure Digital Voice Communications in World War II*, http://www.nsa.gov/wwii/papers/start_of_digital_revolution.htm (June 2003).)

rated [60]. The paper of Halsey and Swaffield [61] from the British Post Office Research Station shows the state of the art in vocoders in the UK in 1948. The paper reports nothing new, apart from details about circuitry. Pitch and voicing detection gets quite some attention and it is discussed to be a delicate process. Analog TDM of the parameter signals has been proposed by Vilbig and Haase of the USA Airforce Cambridge Research center in 1956 [62]. A paper of Gold and Rader from the USA Air Force oriented Lincoln Laboratory of the Massachusetts Institute of Technology (MIT) in 1969 [63] shows that hardly any progress could be reported in the design and performance of channel vocoders. In the 1970s the circuit implementations of channel vocoders became digital [64, 65, 66] but the basic functionality remained the same. The paper of Holmes [67] of the Joint Speech Research Unit in the UK shows that in 1980 still not much progress could be reported. The last step in channel vocoder development was the implementation on ICs, as well on ASIC (Application Specific IC) [68] as on microprocessor [69], and also on the first general purpose digital signal processor μPD7720, put on the market by Nippon Electric Company (NEC) [70, 71]. The relatively

high computing power of these devices enabled the application of more complex algorithms resulting in significantly improved performances of pitch and voicing detection. Pitch and voicing detection are discussed in a separate subsection (Subsection 3.3.4), however, because they have a wider application area than that of channel vocoders alone, including other types of vocoders such as formant and LPC vocoders.

The swan song of the channel vocoder was perhaps the rejection of the channel vocoder with a "harmonic sieve" pitch detector (see Subsection 3.3.4) of a company in the UK [72], in a contest in the early 1980s for a new NATO secure telephone unit (STU) standard. The survivor in this contest was the LPC vocoder which is discussed in Subsection 3.3.3. In the 1980s, the LPC vocoder definitely succeeded the channel vocoder.

3.3.2 Formant Vocoders

The formant vocoder appears for the first time in the literature in 1950, in an abstract [73]. In this abstract, Munson and Montgomery report that work on a resonance type of vocoder was started in the early 1940s, but discontinued during the war.

In their analyser, speech was split into four bands, i.e. 40–400, 300–1000, 1000–3300, and the band above 3000 Hz, see Figure 3.10. In each of the upper three bands the zero-crossing rate was determined as a first approximation to the frequency of the dominant formant in that band. The amplitudes of the formants were estimated on the basis of the envelope of the total signal in the band. Furthermore, a pitch frequency was measured in the 40 - 400 Hz band and another amplitude parameter, representing this band, was determined. The bandwidths of these eight parameters initially used by Munson and Montgomery was 40 Hz, but it could be limited to the same value as in the channel vocoder (25 Hz). In this way, the formant vocoder promised a lower bit rate than the channel vocoder.

In the synthesiser the excitation function is generated by the usual impulse and noise generators, but an improved voicing decision is made on the basis of the balance of the amplitude parameters of the lowest and the highest frequency bands. The upper three bands are regenerated using a variable-frequency resonator in each band, thus reintroducing the estimated formant structure. The resulting band-signals are weighted according to their received amplitude parameters and added together with the reconstructed lower band (< 400 Hz) to complete the synthetic

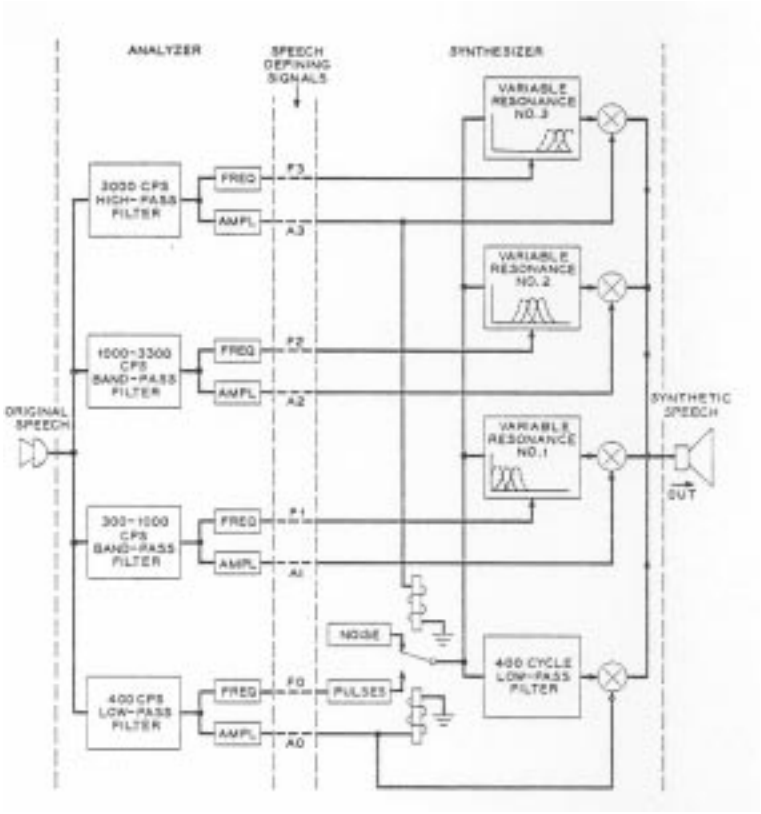

Figure 3.10: Formant vocoder after Munson and Montgomery (from the book of Flanagan [48]).

speech output signal.

The promising properties of the formant vocoder attracted other researchers, amongst whom James Flanagan and Arthur House of the Air Force Cambridge Research Center and Acoustics Laboratory of the MIT. In 1956 they report on a formant-coded speech compression system [74] using only 7 parameters: the pitch F_0, the first three formant frequencies F_1, F_2 and F_3, an amplitude parameter A_v for voiced excitation, another one A_n for noise excitation, and a frequency F_n representing the frequency of the major concentration of speech energy in the 3000-7000 Hz band. This system differs from the original system of Munson and Montgomery in several aspects, as summarised below.

- There is no voicing decision, but instead there are two amplitude

parameters controlling the amplitudes of a pulse excited branch and a noise excited branch which are added together in a way similar to the speech production model of Figure 2.34.

- It uses a synthesiser consisting of a *cascade* of second-order resonators after a terminal analog model (see Section 2.3.7) of the vocal tract, instead of the parallel architecture of Munson and Montgomery.

- A formant is characterised by its frequency location and another parameter. In the Munson and Montgomery system this is the formant-amplitude. In the cascade architecture this is the formant bandwidth. These formant-bandwidth parameters are absent in the proposed system, however. This is based on previous work from 1953 on measurements of formant bandwidths by Bogert [75], from which it was concluded that the formant bandwidths are approximately constant in the frequency range up to 3000 Hz. In 1961, however, Dunn showed that this was quite a simplification [76].

- Formant analysis is based on more sophisticated spectral peak-picking [77], rather than on the crude method of counting zero-crossings.

Later, in 1972, Flanagan wrote in his book [48] in Section 8.5 on formant vocoders that bit rates of 1000 bit/s, and even lower, had been realized, but that the quality remained relatively poor. Elaborate tests on the above system showed that vowel sounds are identified correctly more than 80% of the time, and that consonant sounds are identified correctly only 25% of the time.

The problem of reliable formant analysis and tracking over time has never been solved completely, albeit that the application of digital LPC techniques brought quite some improvements [78]. The digitalization wave of the 1970s also brought digital implementations of the formant vocoder, of course. A good example is the 1200 bit/s digital formant vocoder of Chong Kwan Un described in a paper in 1978 [79]. The speech quality of this vocoder was reported as "reasonably good" and that no other vocoder was known to yield better speech quality at 1200 bit/s. However, since the invention of the LPC vocoder in the late 1960s and its development in the 1970s it became gradually evident that the LPC vocoder provided a much more attractive alternative, despite the fact

that a low bit rate in the order of 1000 bit/s could not yet be reached. The once so promising formant vocoder went out of use.

3.3.3 LPC Vocoders

The LPC methods, described in Section 2.5, which have been invented by Itakura and Saito (PARCOR or autocorrelation method) and Atal and Schroeder (covariance method), were the enabling technologies for LPC vocoders. Itakura and Saito published their invention directly in connection to a vocoder in 1968 [80]. Atal and Hanauer reported on an LPC vocoder in 1971 [81]. The basic architecture of these vocoders is straightforward. In the encoder the speech signal is segmented and from each segment of normally 20 ms, a pitch parameter, a voicing parameter, a gain parameter and typically ten LPC parameters are determined. These parameters are quantized, appropriately encoded, and transmitted. In the decoder an LPC synthesis filter is excited by voiced and unvoiced excitation generators controlled by the pitch and voicing parameters, still very much in the way as in Dudley's channel vocoder. The gain parameter controls the excitation level.

The bit rate of the vocoder of Itakura and Saito amounts to 5400 bit/s using 9 bits for each of the ten reflection coefficients - which they called PARCOR coefficients - and 6 bits for each of the three excitation parameters: a pitch parameter, a sliding voicing parameter indicating the ratio of voiced and unvoiced energy, and a gain. The update rate of the parameters is 50 times per second.

In the vocoder of Atal and Hanauer the pitch parameter uses 6 bits, there is a 1-bit voicing parameter and the gain parameter is represented by a 5-bit code. The twelve LPC parameters are transformed into frequencies and bandwidths of the poles of the synthesis filter because this representation turns out to be less sensitive than the direct-form a-parameters. The pole parameters can adequately be coded by 60 bits. At an update rate of once per 30 ms this adds up to 2400 bit/s. In their paper, they interrelate LPC to the composite acoustic tube and it is suggested that quantization of the cross-sectional areas of the tube sections, instead of the pole parameters, can be an attractive alternative.

In 1973, Haskew, Kelly and McKinney [84] proposed logarithmic quantization of the area ratios of the associated acoustic tube. Stability is easily maintained in this case by preserving positive areas. The performance of 58-bit quantization with nonuniform bit allocation on 12 of such parameters was found to be perceptually equal to quantiza-

tion of root locations with 11 bits for complex roots and 5 bits for real roots. (There are usually a few, say two, real roots, representing possible spectral tilt, see Section 2.5.2.) The log-area ratio quantizer is also computationally much more efficient since root locations do not have to be computed. The area ratios can be determined directly from the reflection coefficients by the use of Equation 2.34. In accordance with this equation we define the log-area ratio LAR by

$$LAR_m = \log \frac{\mathcal{A}_{m-1}}{\mathcal{A}_m} = \log \frac{1 + r_m}{1 - r_m} \ . \tag{3.23}$$

The link to the (cascade) formant vocoder is discussed as well by Atal and Hanauer in their 1971 paper. It is argued that the transfer function of the synthesis filter $1/A(z)$, given by

$$\frac{1}{A(z)} = \frac{1}{1 + \sum_{m=1}^{M} a_m z^{-m}} \ , \tag{3.24}$$

where M is the order of the LPC, can be rewritten as

$$\frac{1}{A(z)} = \frac{C z^M}{\prod_{i=1}^{M} (z - p_i)} \ , \tag{3.25}$$

with C being a constant and p_i the poles of $1/A(z)$. For real valued coefficients a_m there can be real poles in combination with complex conjugate pairs of poles. Each pair of complex conjugate poles $\{p_k, p_k^*\}$ represents a resonator characterised by the partial transfer function

$$\frac{1}{(z - p_k)(z - p_k^*)} \ , \tag{3.26}$$

and it produces a spectral peak in $1/|A(e^{j\omega T})|$, where T stands for the sampling period, in the proximity of the pole frequency, which can be associated with a formant. The radian frequency ω_k of the pole is given by

$$\omega_k = \frac{1}{T} \Im(\ln p_k) \ , \tag{3.27}$$

where $\Im(.)$ stands for the imaginary part, and the associated bandwidth b_k is approximated[2] by

$$b_k = -\frac{2}{T} \Re(\ln p_k) \ , \tag{3.28}$$

[2]The exact value is slightly different, see Appendix D.

where $\Re(.)$ stands for the real part.

In the LPC vocoder, no formant detection is required, however, because the formant information is implied in the set of prediction coefficients and the problem of formant tracking from segment to segment does not exist. In [82] (1974), Markel and Gray elaborated further on the LPC vocoder on the basis of the autocorrelation method and finite word-length effects are studied bringing the LPC vocoder closer to real-time hardware implementation. The development of LPC vocoders in the years following their introduction concerned the quantization process of the LPC coefficients, improvements in the excitation function, and efficient hardware implementations.

The quantization problem to be solved was to find an efficient, perceptually adequate, representation of the coefficients with as few bits as possible which at the same time guaranteed stability of the synthesis filter. Already in 1972, Itakura and Saito [83] used a log spectral distance measure F based on an L_2 norm (square error) to optimize a nonuniform bit allocation to their PARCOR coefficients, according to

$$F = \left\langle \left(20 \log \left| \frac{A_n(e^{j\omega T})}{A'_n(e^{j\omega T})} \right| \right)^2 \right\rangle_{\omega,n} , \tag{3.29}$$

where $A'_n(e^{j\omega T})$ represents the transfer function $A_n(e^{j\omega T})$ of the inverse filter with quantized coefficients, n labels successive segments, and where $\langle . \rangle_{\omega,n}$ stands for averaging over frequency and time, respectively. Stability is easily guaranteed by keeping the magnitude of the quantized PARCOR coefficients below unity. The bit allocation is optimised by minimising F on training data (speech) under the constraint of a fixed total number of bits. In their paper, the result for 10^{th} order LPC is compared to a uniform bit allocation of 6 bits per coefficient and it is concluded that the same subjective quality is obtained with a total of 42 bits for the optimised case. The application of a fixed pre-emphasis (differencing with a one-sample delay) on the input speech signal prior to LPC analysis can even bring a slight improvement.

In 1975 Viswanathan and Makhoul report on a quantization method using a log spectral distance based on the L_1 norm (absolute error) [85]. The mean absolute spectral error $\Delta\mathcal{S}$ they used is given by

$$\Delta\mathcal{S} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| 20 \log \left| \frac{A(e^{j\omega T})}{A'(e^{j\omega T})} \right| \right| d\omega T \ . \tag{3.30}$$

The spectral sensitivity, i.e. $\Delta\mathcal{S}$ as a result of a deviation in the reflection coefficient r_m, given by

$$\frac{\delta\mathcal{S}}{\delta r_m} = \lim_{\Delta r_m \to 0} \left| \frac{\Delta\mathcal{S}}{\Delta r_m} \right| , \tag{3.31}$$

is shown to be a $\cup$-shaped function. It is explained that *uniform sensitivity* is obtained if the sensitivity function is evaluated with respect to the LARs instead of the reflection coefficients. The optimal bit allocation is then derived by minimising the maximum spectral deviation $(\Delta\mathcal{S})_{\max}$ due to quantization, which turns out to reduce to equal quantization steps for all LARs.

In the same paper it is noted that the authors observed that the short-time spectral dynamic range of a speech segment is the most important factor that affects the quantization properties. The quantization properties can be improved by reducing this dynamic range. Using a pre-emphasis of the form $1 + a_1 z^{-1}$ with $a_1 = -\rho_1/\rho_0$ (see Section 2.5.1), which can be compensated at the synthesiser, the spectral dynamic range is reduced by removing the tilt from the spectrum. So, it is similar to increase the order of the LPC by one, and by freeing the LPC from this task it will better approximate the remaining spectral details. Sometimes, as mentioned above, pre-emphasis and de-emphasis are applied with a fixed prediction coefficient according to the long-term average of a_1. Although the performance will decrease in this case, no extra coefficient has to be communicated to the decoder.

In 1977, Gray Jr., Gray and Markel propose the *minimum deviation* method [86]. Whereas Viswanathan and Makhoul use a spectral deviation bound (the maximum value of $\Delta\mathcal{S}$) which is minimized, Gray Jr., Gray and Markel propose the use of an expected or mean spectral deviation bound $\mathcal{E}(\hat{\mathcal{D}})$. Here, $\hat{\mathcal{D}}$ is some convenient upper bound of some form of log spectral distance, which has proven to bear at least some perceptual relevance. The precise formulation of it, for instance whether it should be based on an L_1 or an L_2 norm, does not appear to be very critical. The deviation of this spectral measure due to quantization of the coefficient x is described by (compare to Equation 3.5):

$$\mathcal{E}(\hat{\mathcal{D}}) = \sum_{n=0}^{N-1} \int_{x_n}^{x_{n+1}} \hat{\mathcal{D}}(x, y_n) p_x(x) dx , \tag{3.32}$$

where x_n and x_{n+1} represent two decision levels of an N-level quantizer and y_n the intermediate reconstruction level. The coefficient x is the

reflection coefficient, the LAR, or any other monotonic mapping. The expected deviation measure explicitly introduces the probability density function $p_x(x)$ of each coefficient - which can be approximated by histograms - into the quantizer design. Minimization of $\mathcal{E}(\hat{\mathcal{D}})$ results into nonuniform quantizers.

In 1980, Markel and Gray Jr. report on a comparison of specific implementations of the uniform sensitivity and the minimum deviation quantization methods [87]. They conclude that the distortion level achieved with uniform sensitivity procedures at 42 bits/frame can be achieved with 33 bits/frame using the minimum deviation method. These data concern pre-emphasised speech sampled at 6.5 kHz and LPC order 10. In [88] another comparison between the two methods is made, both implemented at 44 bits/frame with LPC order 12 and sampling frequency 8 kHz and without pre-emphasis. Also here, the results show the superiority of the minimum deviation method.

Another representation of the prediction coefficients which has turned out to be very useful for quantization purposes, are the line spectral frequencies (LSFs). They have been proposed by Itakura in 1975 [89] and he called them line-spectral pairs (LSPs). They are obtained from the polynomials of two extended versions of the (lattice) inverse filter according to Figure 2.28. One version is obtained by extending the M^{th} order inverse filter to order $M+1$ with a reflection coefficient $r_{M+1} = 1$. The other version is obtained by using $r_{M+1} = -1$. In this way the transfer functions of the inverse filters become (see Equation 2.54)

$$A_{M+1}(z) = A_M(z) \pm z^{-(M+1)} A_M(z^{-1}) \; . \tag{3.33}$$

If we denote $A_{M+1}(z)$ with $r_{M+1} = 1$ by $\mathcal{P}(z)$ and in the case of $r_{M+1} = -1$ by $\mathcal{Q}(z)$, then

$$A_M(z) = \frac{1}{2}\{\mathcal{P}(z) + \mathcal{Q}(z)\} \; . \tag{3.34}$$

From Equation 3.33 it follows that $\mathcal{P}(z)$ is a symmetric polynomial and $\mathcal{Q}(z)$ is an antisymmetric polynomial. Important properties of $\mathcal{P}(z)$ and $\mathcal{Q}(z)$ are summarised in the following list.

- If M is even, $\mathcal{P}(z)$ has a real zero for $z = -1$ and the polynomial $\mathcal{P}'(z) = \mathcal{P}(z)/(1+z^{-1})$ is of even order M. In this case $\mathcal{P}'(z)$ has $M/2$ complex conjugate zero-pairs. For even M $\mathcal{Q}(z)$ has a real zero at $z = 1$ and $\mathcal{Q}'(z) = \mathcal{Q}(z)/(1-z^{-1})$ is again of order M and has $M/2$ complex conjugate zero-pairs as well.

- If M is odd, $\mathcal{P}'(z) = \mathcal{P}(z)$ is of even order $M+1$ and has $(M+1)/2$ complex conjugate zero-pairs. In this case $\mathcal{Q}(z)$ has two coinciding zeros at $z = 1$ and $\mathcal{Q}'(z) = \mathcal{Q}(z)/(1-z^{-2})$ has $(M-1)/2$ complex conjugate zero-pairs.
- All zeros of $\mathcal{P}'(z)$ and $\mathcal{Q}'(z)$ are on the unit circle and the (positive) frequencies of the zero-pairs are called LSFs or LSP frequencies.
- The locations of the LSFs of $\mathcal{P}'(z)$ and $\mathcal{Q}'(z)$ on the unit circle are interlaced.
- The minimum phase property of $A(z)$ is preserved if the frequency order of the LSFs is not changed. In the case of quantization of the frequencies, this constraint can easily be met.

The latter three properties have been shown by Soong and Juang in 1984 [90]. Ways to find the zeros of the polynomials have been discussed by them and by Kabal and Ramachandran [91], amongst others. Soong and Juang report in their paper on an objective comparison based on a likelihood distortion measure between a quantization scheme using 30 bits for ten LSFs and another scheme using 43 bits for ten LARs. The results show that with this 30% reduction in bit rate the LSF quantization scheme achieves an even (slightly) lower average distortion level. A year later, Kang and Fransen [92] reported on subjective evaluations. They found that the DRT (diagnostic rhyme test, see Section 1.2.4) of their low bit rate vocoder using ten LSFs and a 31-bit quantization scheme was virtually identical to the case where ten reflection coefficients were quantized using 41 bits.

The impulsive excitation source for voiced sounds was suspected to cause an artifact that seemed to be produced by all LPC vocoder implementations. The particular rattle-like sounding artifact became known as "buzziness" or simply the "buzz" [93]. Proposed remedies included the use of Rosenberg excitation pulses (see Section 2.5.3) instead of the usual Dirac-like pulses, the use of all-pass networks to achieve a similar goal of smearing the excitation energy over time, the application of time jitter to the excitation pulses to break their stiff periodicity, and the use of mixed-source models [94], the basic idea of which has been discussed in Section 2.5.3. Hedelin [95] used variable shape "glottal pulses" which were represented by four parameters each and which were analysed from each particular speech segment. He found that "the resulting speech quality was far above that of standard vocoders". A drawback of the method was that it was "computationally very intensive".

In the years around 1980 the viability of LPC vocoders was proven by early hardware implementations for use in full-duplex telephony.

In 1979 an LPC vocoder at 2400 bits/s, called TSP-100, is introduced to the market by Time and Space Processing Inc. [96]. The vocoder works with 22.5-ms frames each producing 54 transmission bits. The particular LPC-method used is based on 9-th order lattice filters. For each of the first two reflection coefficients, 6 bits uniform quantization is used, 5 bits uniform quantization for the third reflection coefficient and 4 bit quantization for each of the 6 remaining coefficients in the form of LARs. The pitch and voicing detection make use of the AMDF technique, to be discussed in Subsection 3.3.4, and the pitch is represented by a 7 bit code. Zero value of this code indicates the unvoiced case. A 6 bit parameter controls the signal level. The intended applications are encryption to provide security over ordinary voice-grade telephone links, multiplexing of four voice channels on high-grade links to save user costs, and the ability to multiplex speech with other digital data. Multiplexing offers an economic advantage, especially in PBX environments, according to [96]. The complete codec is realized in a desktop unit containing four printed circuit boards. Two microprocessors 8080-1 are incorporated, one for the analyser and one for the synthesiser. Computationally intensive procedures are executed by special purpose hardware. The system has a power consumption of almost 40 W.

In 1981, we realized an own LPC vocoder [97], at a bit rate of 2400 bit/s. Every 22.5 a frame was analysed resulting into 54 transmission bits. The ten LPC parameters were represented in the form of LARs and quantized into 41 bits. The pitch was determined from the input speech signal on the basis of our own "harmonic sieve" method, see Subsection 3.3.4, and quantized to a 6-bit code including binary voicing information. A 5-bit gain parameter was used for level control and the last two bits in the frame were reserved for frame synchronisation purposes. The synthesis filter in the decoder was of the lattice type. Its hardware was built around Signetics 2901 microprocessor chips and the legendary TRW parallel multiplier was used to speed up computations.

The most representative LPC-vocoder implementation of the time was the 2400 bit/s USA Government - Department of Defense (DoD) - Standard LPC-10 vocoder from 1982 [98], and for many years thereafter, it has worldwide been used as *the* reference vocoder. It was intended to be used in combination with digital encryption techniques to provide secure voice, as a modern successor of Sigsaly. This vocoder also used

	Natural Speech	LPC-10
DRT (% correct)	about 95	90
DRT with noise (% correct)	92-93	above 82
DAM (quality score)	about 65	48

Table 3.2: Speech quality of LPC-10 as compared to natural speech (after [98]).

22.5-ms frames of 54 bits each. Pitch was measured using the AMDF method, see Subsection 3.3.4, and four voicing states were distinguished, i.e. steady state voiced, unvoiced to voiced transition, voiced to unvoiced transition and steady state unvoiced. Pitch and voicing were coded by 7 bits and a gain parameter took 5 bits. For frame synchronisation only 1 bit per frame was used. For voiced frames the LPC coefficients were non-uniformly quantized using 41 bits. For non-voiced (unvoiced and transition) frames only fourth-order LPC was used taking 20 bits, leaving 21 bits for error protection. In the decoder pitch smoothing techniques were applied, which are discussed in Subsection 3.3.4, and synthesis was realized on the basis of a direct-form recursive filter. The voiced excitation pulses were time-spread by the application of an all-pass filter. In the speech input path a first-order pre-emphasis was used to reduce the required accuracy in the LPC analysis, and a corresponding de-emphasis was applied to the speech output in the decoder. The performance of this vocoder, as compared to natural speech, is shown in Table 3.2. The DRT component measures used were the major phonetic attributes of speech: voicing, nasality, sustention, sibilation, graveness, and compactness. The noise environments used included the typical office, airborne command post, ship, helicopter carrier, jeep and tank. In the DAM (diagnostic acceptability measure, see Section 1.2.4) test two component measures were used: a measure of the quality of the perceived background noise and the perceived signal sound.

Another interesting LPC vocoder implementation has been realized at the MIT Lincoln Laboratory in Lexington in 1983 [99]. This one was based on the first single-chip signal processor μPD7720 of Nippon Electric Company (NEC) [100], which had been introduced to the market in 1980, and it was controlled by an Intel 8085 microprocessor. For the complete vocoder, a total of 16 integrated circuits was used occupying 18 square inch of circuit area. The power dissipation was only 5.5 W.

3.3.4 Pitch and Voicing Detection

In 1937 a journal paper by the German scientists Grützmacher and Lottermoser on recording of melody curves of speech and chant appeared in [101]. This paper was not the first one on this subject. The subject was already discussed in the literature in the late 1920s. In those early days the measurements were often based on a bank of (mechanical) resonators in which neighbouring resonators were tuned only a few Hertz apart. The method of Grützmacher and Lottermoser used electronic signal processing techniques to measure the melody. A quadratic characteristic was applied to enhance the amplitude of the difference frequency between the harmonics of the speech signal and a subsequent low-pass filter had the task to further boost the amplitude of this fundamental over the higher harmonics. Then, a frequency meter on the basis of measuring the time duration between successive zero-crossings of this signal generated a signal proportional to this pitch.

In the vocoder of Dudley, described in Section 2.3.2, a similar method was used as appears from his 1939 publication [102]. It is very instructive to quote the passage on pitch detection from this paper, showing the state of the art of the time.

> "The pitch analysis consists in obtaining the fundamental frequency of the voice in reasonably pure form and then measuring it with a frequency meter.
>
> The purification of the fundamental frequency results from a frequency discriminator which cuts out both the higher frequencies of the voice and the very low frequencies, particular as the frequency becomes less than 50 cycles so as to eliminate certain puffs from the voice which are of large enough power to be troublesome in pitch analysis. The loss at the higher frequencies is made sufficiently sharp to insure that the fundamental frequency itself will get through much the strongest of any component and so be in reasonably pure form (...).
>
> The reasonably pure fundamental frequency obtained from the frequency discriminator is applied next to a frequency meter circuit (...) This circuit sets up a fairly uniform pulse each time the current swings to a sufficiently positive or negative value (...) The result is a number of pulses proportional to the frequency measured. If the upper harmonics of the voice are not sufficiently eliminated by the discriminator

> then more than two pulses are received per cycle of fundamental frequency giving the effect of increased pitch. When this happens occasionally, the circuit swings from normal pitch measurement to an apparently higher frequency pitch measurement leading to an unpleasant raucousness in the remade speech.
>
> The output pulses from the frequency meter are next passed through a 25-cycle low-pass filter (...) This low-pass filter serves to eliminate all the original frequencies of the voice but to pass a current proportional to the number of pulses received and therefore approximately proportional to the original fundamental frequency. At the output of the low-pass filter, then, there is a pitch-defining current, thus completing the pitch analysis."

Unvoiced sounds are supposed not to provide enough power to set up pulses in the frequency meter and zero output will result. In this way, an implicit voiced/unvoiced decision was obtained. Dudley did not use quadratic or non-linear characteristics, but purification of the fundamental was obtained by linear filtering. This required special microphones which were able to pick up the low pitch frequencies down to 50 Hz. The method did not work for telephone speech with a pitch below 200 Hz or so, because the fundamental itself is not contained in such speech. Furthermore, Dudley reports that occasionally erroneous pitch measurements occur.

In fact, the problem of pitch detection in a speech signal is the search for periodicity in a signal which is at best "almost" periodic for a relatively short period of time, because the pitch, and also the spectral characteristics, are continuously varying. Sometimes the pitch periodicity is even weaker than the periodicity induced by a strong formant, normally the first formant, thus masking the true pitch and putting forward a higher pitch harmonic, as Dudley already reported. On the other hand, it also occurs that sequences of two or even more pitch periods show a stronger periodicity than the true pitch. Another source of indefiniteness are the transitions from unvoiced to voiced and vice versa. Pitch detectors suffer from these opacities and they make errors.

In 1949 Gruenz and Schott discuss in a paper [103] the extraction and "portrayal" of pitch of speech sounds. Several innovations were proposed. To start with, they recognized that it is generally necessary to handle pitch frequencies up to 600 Hz, while Dudley's implementation

could only handle pitch frequencies up to half this frequency. Moreover, an explicit voiced/unvoiced detector was used, based on the balance of energies from the 150-900 Hz band and the band above 4500 Hz, respectively. For voiced sounds the energy in the low-frequency band with its strong first formant will prevail, while a paramount presence of the high-frequency band only occurs in unvoiced sounds. Furthermore, gain control was applied to handle a large range (25 dB) of amplitude variations and "double detection" was used as alternative purification of the waveform. Double detection uses an envelope detector incorporating half-wave rectification to follow the pitch induced signal envelope variations (see for instance Figure 2.3), a subsequent differentiator to remove the dc component and a second envelope detector. The intended result is that by non-linear processing the complex speech waveform is transformed into a simple waveform with only one outstanding peak per pitch period. Then, a frequency meter can readily generate a signal proportional to the peak frequency. They report that "reliable indications of pitch have been obtained over a range corresponding to frequencies from 100 to 600 cycles for a wide range of voices".

In 1962, Bernard Gold of the MIT Lincoln Laboratory proposed a computer program for pitch extraction [104]. Among the existing methods of pitch detection, he distinguished between two basic methods. One method uses linear filtering to enhance the (quasi-)periodic fundamental and the pitch is found on the basis of what he called "regularity of large sections of the speech wave", by which he meant periodicity. The other method makes use of features in the waveform, essentially of the "peakedness" of a single pitch period, and non-linear processing is used to generate a pulse train of pitch markers. In his proposed computer program he combined the two methods and expected that the two methods would work in a complementary way so as to outperform both basic methods. Essentially, the method boiled down to three parallel pitch detectors each observing different features of the waveform which were then evaluated for periodicity, rather than the waveform itself. The final decision on the position of consistent pitch markers was made on the basis of those evaluations. Gold concluded that "Although our program is not as good as a man's eye in selecting the markers, it is not far removed from this ideal." Two years later he completed his work by presenting a new voiced/unvoiced detector, based on the presence or absence of periodicity [105].

In 1969 his pitch detection strategy had been developed to the use of

six relatively simple parallel pitch detectors, each detecting peaks and valleys in low-pass filtered speech (with a cut-off frequency of 600 to 900 Hz), and the application of sophisticated coincidence measurements as part of the final pitch decision based on appropriate majority logic. This was published in a paper together with Rabiner [106]. Voicing detection based on the presence of sufficient periodicity was included. This method became known as the "parallel processing" method and it has widely been recognized as a reference pitch and voicing detector for many years.

In the mean time methods based on the autocorrelation function were gaining more and more interest. A forerunner, based on determining the difference

$$d(t,\tau) = |s(t) - s(t-\tau)| \,, \tag{3.35}$$

was proposed by Miller and Weibel already in 1956 [107], see also [112]. The speech signal $s(t)$ is delayed - using a tapped delay line - over a time τ and subtracted from $s(t)$. For an exact periodic signal this results in zero output if the delay τ is adjusted to one period (but also to multiples of the period). Unvoiced sounds can be detected by the absence of such nulls. The short-term autocorrelation function, however, is given by

$$\rho(\tau, t_0, t_1) = \int_{t_0}^{t_1} s(t)s(t+\tau)dt \,. \tag{3.36}$$

This function peaks at $\tau = 0$ and at values of τ equal to multiples of the pitch period. Generally speaking, the time limits should be set to $t_0 = -\infty$ and $t_1 = \infty$. In this case $\rho(\tau)$ has become independent of t_0 and t_1 and it is a symmetric function with respect to $\tau = 0$, so using a delayed signal $s(t-\tau)$ instead of $s(t+\tau)$ gives the same result. For short-time observations the signal should be windowed to limit its time duration with zero signal outside this interval. The windowed signal is then shifted in time to evaluate the short-time autocorrelation function. The window length should be set to contain at least more than one period for the lowest expected pitch frequency. Consequently, the integration interval $t_1 - t_0$ can be shortened to include at least all non-zero contributions to the integral. The discrete-time autocorrelation function is accordingly defined by

$$\rho[m] = \sum_{n=n_0}^{n_1} s[n]s[n+m] \,. \tag{3.37}$$

Already in 1959, an early implementation is reported by Gill [108]. He used binary circulating delay lines containing 16.2-ms segments of clipped speech sampled at 5.5 kHz and the required multiplication for the correlation was realized by a modulo-2 adder. The pitch is found by looking for the $m \neq 0$ that gives the highest correlation peak. The amplitude of the peak is used as a measure for voicing, based on the fact that a high correlation peak indicates a strong periodicity and vice versa. The only performance indication given by Gill is that "gross errors of pitch are rare". Figure 3.11 shows an example of a (normalised) autocorrelation function according to Equation 3.37 of the speech segment of Figure 2.3. The normalised function is obtained by dividing the function by its value at zero lag so that it takes the value one there. The peak at a lag of 90 sampling periods (8 kHz sampling frequency) represents the pitch period and the peaks spaced approximately 22 sampling periods apart show the interaction of the formant which occurs at about 365 Hz in Figure 2.4. The horizontal line at $\rho[m]/\rho[0] = 0.2$ is a possible threshold value for voiced/unvoiced discrimination. The first formant-induced peak at a lag of about 22 sampling periods shows that if the pitch was defined as the first peak above the threshold, a pitch measurement error would have been made.

The autocorrelation method came to maturity by the work of Man Mohan Sondhi of Bell Telephone Laboratories. He realized that formant interaction in the autocorrelation function could be removed by the use of spectral flattening. This approach is mentioned for the first time in a conference paper in 1965 [109]. In a paper in 1968 he presents three new methods of pitch extraction: spectrum flattening and subsequent phase synchronisation of the harmonics, spectrum flattening followed by autocorrelation, and center clipping (for efficient spectrum flattening) followed by autocorrelation. In the introduction of this paper he wrote:

> "The basic notion common to all three pitch extractors described here is the following. If the harmonics of the fundamental frequency could be made equal in amplitude and put into phase synchronism with each other, the resulting waveform would be a train of highly peaked pulses, and the interval between these pulses would correspond to the current pitch period. During unvoiced intervals, no such pulse train would be obtained and, thus, a v/uv decision could be based upon the presence or absence of the pulse train."

Figure 3.12 shows the normalised autocorrelation function of the

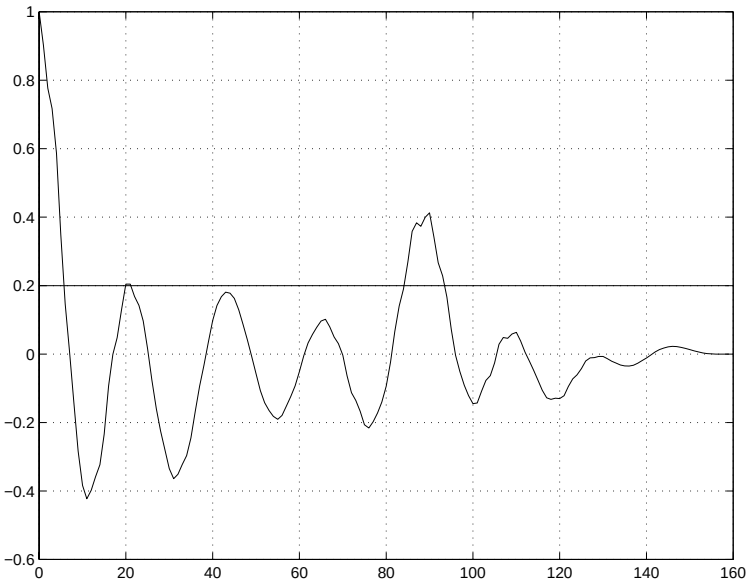

Figure 3.11: Normalised autocorrelation function. The horizontal axis shows the lag m in 8 kHz sampling periods and the vertical axis shows $\rho[m]/\rho[0]$. The horizontal line at $\rho[m]/\rho[0] = 0.2$ is a possible threshold value for voiced/unvoiced discrimination.

same speech segment as used in Figure 3.11, but now spectrally flattened. The spectral flattening is realized by a tenth-order LPC inverse filter. Sondhi's message is clearly demonstrated by this picture. The use of the same threshold value as in the previous picture (0.2) does not suffer from any formant interaction, anymore. In addition, the location of the correlation peak leaves no room for any doubt, there is only one outstanding sample.

If it is true that a better compliance with these impulse-forming conditions also gives a better performing pitch detector, then we have an objective measure to predict the performance of pitch detectors. From now on, we will refer to this theorem as "*the theorem of Sondhi*":

> A better compliance of the input signal of a pitch detector with the impulse-forming conditions gives a better performing pitch detector.

Indeed, the performance of the latter two pitch detectors, equipped with sophisticated peak detectors, is found to be highly reliable, without

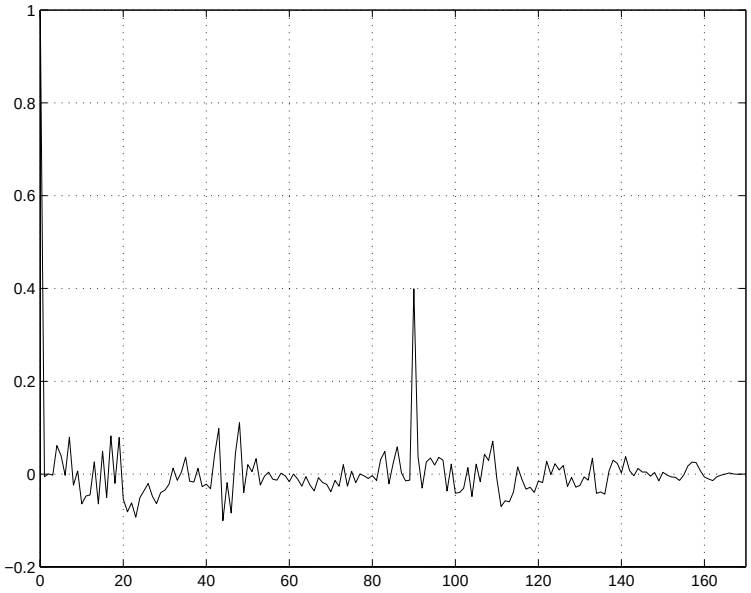

Figure 3.12: Normalised autocorrelation function of the same, but now spectrally flattened, speech segment.

the usual errors of pitch doubling and loss of crisp voiced onsets. The first method is further discarded because it was based on a minimum phase assumption and corresponding phase corrections on individual components appeared too far fetched. On the basis of the theorem the good performance of the other two methods could be expected since they combine spectral flattening and zero-phase setting of all components. The autocorrelation function is equivalent to the inverse Fourier transform of the power spectrum, which is a real, and therefore essentially a zero-phase spectrum. In addition, high-pass filtering of the input speech according to telephone channels and the addition of white noise at an SNR of about 18 dB, are easily tolerated according to Sondhi. In the inverse Fourier transform, white noise maps to a peak at zero lag in the autocorrelation function, and thus does not affect correlation peaks at higher lag values.

Remarkably, the 1968-paper of Itakura and Saito [80], in which they present the first LPC vocoder as discussed in Section 3.3.3, also describes a pitch detector which consists of autocorrelation of the spectrally flattened speech by using the inverse filter $A(z)$. According to Sondhi's

theorem, this was the right way to go. The authors conclude that the advantages of this pitch detector were similar to the cepstrum pitch determination in many respects, except that this method was carried out in the time domain.

The real (the complex form also exists) cepstrum $c[n]$ is defined by

$$c[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \log |S(e^{j\theta})| e^{jn\theta} d\theta \tag{3.38}$$

in which $\theta = \omega T$ is the normalised radian frequency with respect to the sampling frequency $1/T$ [110] (1964). In Equation 3.38 $S(e^{j\theta})$ stands for the discrete-time Fourier transform of a finite duration segment. The cepstrum is an autocorrelation-like function in the sense that if the spectrum is squared and the logarithm removed, the autocorrelation function is obtained, i.e. the inverse Fourier transform of the power spectrum of a segment of the speech signal $s[n]$. The length of the segment is chosen in the same way as for the autocorrelation method discussed above. The cepstrum shows a peak at one pitch period distance from the origin $n = 0$ (and multiples thereof). Detection of the peak position yields the pitch period. So, it can be concluded that the method works with a zero-phase function and the amplitude spectrum is flattened according to a logarithmic compression function. In practice, this appears to be an effective spectral flattener, so that the method practically complies with Sondhi's theorem. If the same peak picking algorithms would be applied, then the autocorrelation and cepstrum methods should give comparable results. This agrees with the pitch detector performance statement of Itakura and Saito in their 1968-paper.

The simplified inverse filter tracking (SIFT) algorithm of John Markel from 1972 [111] is another example of a pitch detector which complies with Sondhi's theorem. The SIFT algorithm is basically not different from what Itakura and Saito did. The speech signal is spectrally flattened by an LPC inverse filter $A(z)$. On the residual signal autocorrelation is applied and the peak indicating the pitch is detected. Voicing decisions are made on the basis of the relative amplitude (with respect to the zero-lag peak) of the pitch peak, as usual. The difference with the other already existing implementations of the same idea lies in the details. The SIFT algorithm works on low-pass filtered speech with a cut-off frequency of 800 Hz. This is a good idea because the higher frequencies - say above 1 to 1.5 kHz - often do not contribute to, or even hamper, a reliable pitch measurement, since the desired stationarity

of speech within a segment usually strongly decreases with frequency. Furthermore, the details of the peak detector and the heuristics applied to create a consistent pitch track over time, can differ greatly, and have their effect on the performance of the pitch and voicing detection. The conclusion of Markel is that the performance of his implementation shows no significant differences with the performance of the cepstrum method. This is precisely what one would expect on the basis of Sondhi's theorem.

Another method from that time (1974) is based on the average magnitude difference function (AMDF) [113], defined by

$$AMDF[m] = \frac{1}{N} \sum_{n=1}^{N} |s[n] - s[n-m]|, \quad m = 0, 1, 2, \ldots, M \ , \qquad (3.39)$$

in which N is the length of the segment and M is the maximum pitch period of the range, usually 16 to 20 ms. It is essentially the same method as the method of Miller and Weibel which had been proposed 18 years earlier, but the difference is again in the particular implementation. The AMDF is implemented digitally and it therefore allows for sophisticated peak detection, tracking heuristics and voicing logic. On the other hand it does not comply with Sondhi's theorem since no spectral flattening was incorporated. The results, as reported, were in "good agreement, differing mainly at the onset or trailing of voicing", when compared to the autocorrelation method. Which autocorrelation method was not precisely defined, though. In 1979, Un and Yang report on an AMDF pitch extractor which is preceded by LPC inverse filtering [114], and it is stated that pitch errors due to formant interaction are avoided.

In 1976, an objective comparison of specific implementations of 7 different pitch detection methods is reported [115]. These methods include the autocorrelation, cepstrum, SIFT, Gold's and Rabiner's parallel processing method (indicated by PPROC), and AMDF. Voicing decisions are based on the previously mentioned measurements of relative signal power, the presence of periodicity, the relative magnitude of the autocorrelation peak or a related other peak as in AMDF or cepstrum pitch detectors, the energy balance between a low-frequency band and a high-frequency band and a related time-domain alternative which counts the number of zero-crossings in the speech segment. The comparison was made for several speech conditions and for several speakers. The results showed no strong overall preference for a particular pitch and voicing detector (-implementation). In a subsequent subjective performance eval-

uation of the same pitch detectors [116] when all applied in the same LPC vocoder, it is concluded that differences in mean preference scores across these detectors were fairly small. The mean preference scores ranged from 1 to 9. The highest preference score (9) was uniformly assigned to the natural speech which was included in the test set. The next highest preference score (6.5) was assigned to the vocoder speech with "ideal", hand edited, pitch and voicing parameters. This gives a good indication of the quality level of the pitch-excited LPC vocoder itself. Using the pitch detectors under test, scores between 3 and 5 were obtained. These preference scores were obtained by averaging across six speakers, three recording conditions, four sentences, and eight listeners.

The last pitch detection method we include in our discussion is the "harmonic sieve" method [117, 118, 119, 120, 121, 122]. Rather than looking at pitch detection as a purely technical problem, the harmonic sieve technique is based on a model of pitch perception. The main functional ingredients are spectral analysis according to the human hearing system, the extraction of relevant component frequencies from the spectrum and subsequent harmonic pattern recognition. The frequency analysis is realized on the basis of the absolute value of the discrete-time Fourier transform. This amplitude spectrum is then subject to a components search procedure, which detects local maxima to find the first N (typically $N \leq 6-8$) relevant components starting at the low frequency end. In this way a frequency set $\{x\} = \{x_1, x_2, ..., x_N\}$ is found which is subject to the harmonic pattern recognition in the form of the harmonic sieve. Note that after arriving at this frequency set, amplitude and phase information is discarded so that compliance with Sondhi's theorem could be claimed, at this stage. The harmonic sieve is a one-dimensional sieve in the frequency domain. A set of I harmonic sieves is considered in which the i^{th} sieve ($i = 1, 2, ..., I$) has a fundamental frequency S_i and J meshes at harmonic frequencies $f_i[j] = jS_i$, $j = 1, 2, ..., J$ (see Figure 3.13). The range of S_i covers the range of pitch in speech, i.e. 50 - 500 Hz in 40 steps, each of one semitone (a semitone corresponds to a factor $2^{1/12}$ in frequency). In accordance with the frequency dependent resolution of the auditory system the mesh at frequency $f_i[j]$ has a relative bandwidth of $2b$, i.e. it extends from $(1-b)f_i[j]$ to $(1+b)f_i[j]$. In this way, all meshes cover equal intervals on a logarithmic scale. The value of b amounts typically to a few percent. Sifting set $\{x\}$ by the i^{th} sieve results in a number of components that pass. These components are labelled with the harmonic numbers j of the meshes through which

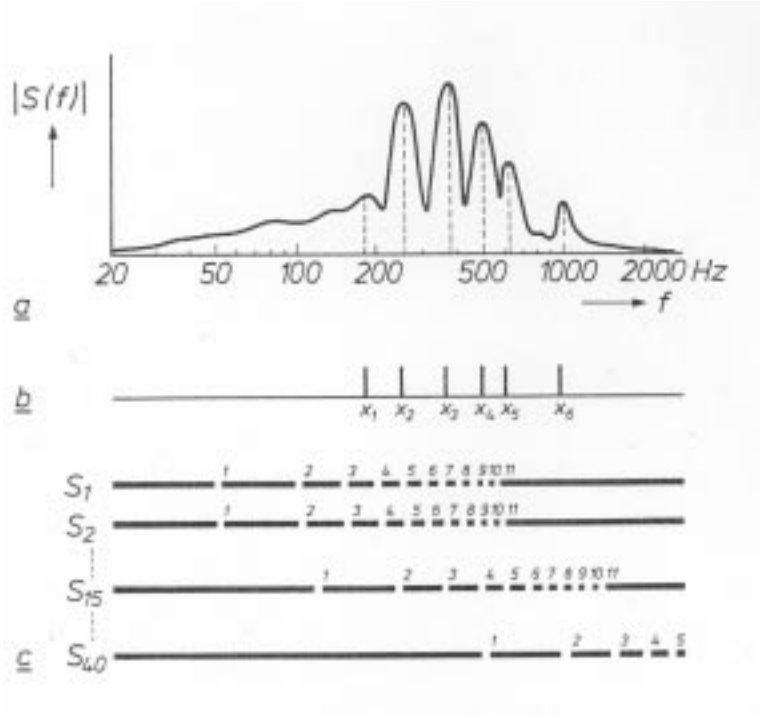

Figure 3.13: Harmonic-sieve pitch detector. From the amplitude spectrum $|S(f)|$ (a) a set of components $\{x\}$ (b) is extracted which are subject to the sieve procedure with 40 successive fundamental frequencies $S_1, S_2, \ldots, S_{40}$ (c).

they pass. The others are labelled zero. This yields a set of harmonic numbers $\{k_i\} = \{k_i[1], k_i[2], ..., k_i[N]\}$ according to

$$k_i[n] = \begin{cases} j & if\ (1-b)f_i[j] < x_n < (1+b)f_i[j] \\ 0 & \text{otherwise} \end{cases} , \tag{3.40}$$

and where n is any integer number between 1 and N. By applying all I harmonic sieves successively to the set $\{x\}$, I sets of harmonic numbers $\{k_i\}, i = 1, 2, ..., I$ are obtained from which the set $\{\hat{k}_i\}$ is chosen which is optimum for $\{x\}$ according to some criterion. In [118], [119], [121] and [122] several criteria are discussed including the Euclidean distance between $\{x\}$ and a candidate $\{k_i\}S_i$ both represented as vectors in a multi-dimensional space. Once the harmonic numbers have been assigned, a refined pitch frequency f_0 is determined on the basis of a

minimum mean-square error between $\{x\}$ and $f_0\{\hat{k}_i\}$:

$$f_0 = \frac{\sum_{n=1}^{N} x_n \hat{k}_i[n]}{\sum_{n=1}^{N} \hat{k}_i^2[n]} \,. \tag{3.41}$$

The performance of the pitch detector is discussed in [122] and it is also compared to the parallel processing method PPROC. One of the conclusions is that for the particular test material the parallel processing method makes octave errors in 2.4% of the cases while the harmonic sieve pitch detector without any tracking provisions makes octave error in 2.0% of the cases. In a subjective comparison where the pitch extractors are used in an LPC vocoder, 68% of the listeners preferred the harmonic sieve, 25% preferred the parallel processing method and 7% had no preference.

In 1982, the harmonic-sieve framework was completed by a suitable voicing detector based on tracking of spectral intensity variations which can be incorporated in the pitch extractor with very little overhead [122]. The unvoiced-to-voiced decision is triggered by a fast enough intensity rise and the voiced-to-unvoiced transition is triggered by a certain intensity decrease regardless its speed.

Some pitch detectors, such as the AMDF methods described in [113] and [114], already make use of tracking and smoothing techniques. In general, linear smoothing (filtering) and non-linear smoothing such as median filtering of the pitch contour as well as the combination of them, can remove errors [124]. If these techniques are applied at the receiving side of a communication system, they can help in removing the influence of transmission errors on the pitch contour, as well. Tracking techniques are often part of the pitch detector itself where they are used to influence the decision making of the pitch measurement. An obvious tracker is based on previous measurements. The search space of the current pitch measurement is narrowed down around previous measurements. This, however, includes the possibility that the tracker persists in following a wrong pitch contour. More intelligent trackers are based on dynamic programming in which future measurements are awaited and hard decisions are postponed until different parallel tracks can be evaluated [125, 126]. This requires that the application can afford the introduced delay. Real-time full duplex speech communication applications often cannot.

An elaborate overview of all pitch and voicing detection and post processing techniques until 1983 is made by Hess [123].

3.3.5 Voice Excited Vocoders

The pitch and voicing detection problem led to alternative vocoder architectures that circumvented the problem. The first impulse into that direction was given by Feldman who patented a baseband vocoder in 1954 [127]. In that proposal, the lower half of the telephone band is transmitted in its natural form. The upper half of the band is compressed on the basis of a constrained channel vocoder architecture, in the sense that the synthesiser is excited exclusively by noise.

A significant improvement of this system was introduced by Schroeder and David in 1960 [128]. They showed that proper vocal excitation of the synthesiser can be obtained from a narrow bandwidth part of the original speech. In first instance they investigated a system with a 3.2 kHz baseband and six vocoder channels to cover the band from 3.2 kHz up to 10 kHz. The synthesiser was excited by the *non-linearly distorted* baseband, which generated pitch harmonics in the case of a voiced baseband and wide-band noise from an unvoiced baseband, both covering the entire upper-band. The non-linear circuit they used had a "W"-shaped input/output characteristic which generates four zero-crossings in its output waveform for every zero-crossing in the baseband signal. The authors reported that, in addition to a considerable gain in intelligibility of the vocoder as compared to the baseband alone, the vocoder improved the reproduction quality markedly. Continued work on this system revealed that a baseband of a few hundred Hz in width is sufficient, and that the method is applicable in general to vocoder systems. In this way a compression factor of two to four became possible with a speech quality close to toll quality. Figure 3.14 shows a block diagram

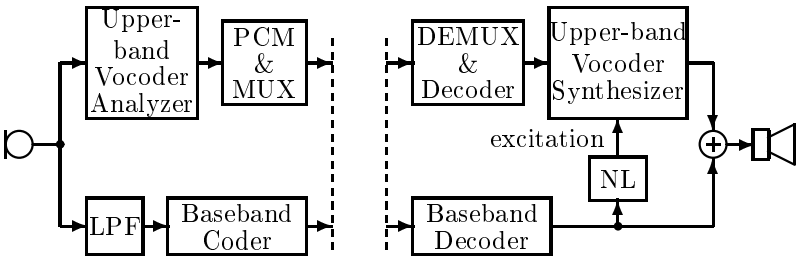

Figure 3.14: Block diagram of a voice-excited vocoder.

of a such a "voice-excited" vocoder. From the input speech the baseband is obtained by a low-pass filter (or possibly a band-pass filter so that the frequencies below 200 or 300 Hz are removed). In the case of analog transmission the baseband signal is transmitted directly to the receiver. In the case of digital transmission the baseband has its own waveform coder and decoder including bit-rate efficient down-sampling in the coder and up-sampling in the decoder. So, the natural baseband signal is part of the output signal and it thus enhances the output speech quality. In the upper-band any vocoder system can be applied and in the case of digital transmission, the parameters resulting from the analyser are quantized, for example by PCM, multiplexed, and aggregated to the digital baseband signal. In the decoder these signals are de-multiplexed and decoded and used in the speech synthesis. The upper-band synthesiser is excited by the decoded baseband signal via a non-linear network (NL) and its output is added to the baseband signal to form the speech output of the voice-excited vocoder.

The principle of voice excitation has been applied to the formant vocoder, as well. Flanagan, the principal advocate of the formant vocoder, see Section 3.3.2, was struggling with the fact that the level of intelligibility of most formant vocoders was significantly below that of the channel vocoder, and he reported on an effort to equip a formant vocoder with voice excitation, already in 1960 [129]. He used a baseband of about 300 to 650 Hz and three formants to cover the telephone band. The total transmission bandwidth was about 600 Hz so that the compression as compared to the conventional telephone channel was about five to six. The non-linear circuit he used generated pulses at the positive going zero-crossings of the baseband signal. Their pitch remained that of the speech itself, for voiced sounds, and their spectrum was sufficiently uniform, according to him. For unvoiced sounds the randomly produced pulses sounded very much like "thermal noise". The obtained performance was described as an appreciably higher intelligibility than that of pitch-excited formant vocoders, but not as good as conventional telephone quality.

The first telephone-band voice-excited channel vocoder was presented by David, Schroeder, Logan and Prestigiacomo in 1962 [130]. They used a baseband from 250 to 970 Hz and 17 vocoder channels to cover the band of 970 - 3700 Hz. The total transmission bandwidth was less than 1200 Hz, thus yielding a factor of three bandwidth compression. The upper-band excitation signal was generated by rectification of the base-

band signal. Special attention had been payed to the accuracy of the voice excitation signal. The spectral envelope of it should be as flat as possible, since all spectral colouring is represented by the channel amplitude parameters. Accordingly, explicit flattening was implemented using a filter-bank to generate narrow bands of the excitation signal and a clipper[3] in each band. Subjective evaluation of this vocoder indicated that, though there was a measurable degradation with respect to the original speech, it was clearly superior to conventional pitch-excited vocoders.

One of the first digital computer simulations of a voice-excited channel vocoder was reported by Golden in 1963 [131]. It concerned a vocoder with a baseband of 250-925 Hz and 10 channels covering the upper band until 3230 Hz. The upper band excitation is obtained by half-wave rectification and differencing, and subsequent spectral flattening using an additional filter-bank and clippers. A completely digital implementation is described, but without the quantization and coding of the parameters to fit low bit rate digital transmission. The software implementation allowed quick evaluation of different parameter settings, especially regarding band-pass and low-pass filter designs. Simulations on an IBM 7090 digital computer required 172 seconds to analyse and resynthesise 1 second of speech. The speech quality is reported to be excellent and often could not be distinguished from appropriately band limited original speech.

A voice-excited channel vocoder for telephone speech with digital baseband transmission was designed and evaluated by Gold and Tierney in 1965 [132]. It was argued for the baseband that the cut-off frequencies should have a ratio of three. A passband ranging from f to $3f$ passes at least two harmonics for fundamental frequencies up to f, since this band is $2f$ wide. For frequencies from f to $3f$ the band contains the fundamental itself. In this way it is ensured that the original pitch up to a frequency of $3f$ is always preserved in the baseband. Because frequencies below 300 Hz are usually absent in telephone speech, they used a baseband of 300 to 900 Hz. By suitable efficient band-pass sampling, a sampling frequency of 1200 Hz sufficed. In combination with ten analog vocoder channels covering the band from 900 to 3300 Hz, 6-bits were required for the logarithmic baseband quantizer to maintain

[3] In the literature the word "clipper" is used but the functionality is more like a "slicer" which gives a predetermined positive output level when its input is positive and for negative input values it generates the opposite output.

the full-band speech quality of the un-quantized case. This amounts to a bit rate of 7200 bit/s for the baseband. Assuming 4 bit per sample and a sampling rate of 50 Hz for the vocoder channels a total bit rate of 9400 bit/s is obtained including a common 4-bit spectrum scale factor. This was the first realistic estimation of the required bit rate for a good quality voice-excited channel vocoder.

Eleven years later, another voice-excited channel vocoder at 9600 bit/s was presented by Zurcher, Graillot and Cartier [133], using a backward adaptive sigma-delta modulation baseband coder at 7200 bit/s. The baseband ranged from 300 to 800 Hz and the upper-band from 800 to 3400 Hz using 8 vocoder channels. In the decoder a simple clipper took care of the high-frequency regeneration. The quality was still at the same level of the previous voice-excited vocoders. Other properties reported were

- a good compromise between complexity and bit rate,
- good speaker recognisability,
- insensitivity to input-signal distortions, and
- high immunity to transmission errors.

This last property is due to the presence of the baseband and its behaviour in the presence of transmission errors. Because the baseband uses the major part of the bit rate, most transmission errors will occur in the baseband signal.

In the mid 1970s we made our own implementation of a voice-excited formant vocoder at a bit rate of 4800 bit/s [134], [135] and [136]. A baseband of 300 - 800 Hz (3 dB) was used and it was sampled at a frequency of 1400 Hz which was determined by the 30 dB bandwidth of about 200 - 900 Hz. The sharp filtering was accomplished by RC-active Cauer filters [137]. In the 800 - 2000 Hz band a single formant with a fixed bandwidth of 100 Hz is allocated, and in the 2000 - 3200 Hz band a fixed formant frequency of 2500 Hz is used with a fixed bandwidth of 200 Hz. Information of the second band which is conveyed to the receiver comprises the formant frequency and the formant amplitude, both sampled at 50 Hz. For the third band only an amplitude parameter is transmitted, also sampled at 50 Hz. The baseband is coded with 3-bit backward adaptive PCM and the formant parameters are each linearly quantized to 4 bits, adding up to a total bit rate of 4800 bit/s. In the synthesis the second formant is excited with the distorted baseband

signal, using just a clipper. The third formant is excited by a white noise source. The speech quality of this vocoder did certainly not meet toll quality, but the sentence intelligibility was good enough to enable a normal conversation. The sensitivity for transmission errors was quite low. At a random bit error rate of 10^{-3} in the transmission bit stream hardly any influence is noticed. At an error rate of 3.10^{-3} disturbances are noticeable, especially during speech pauses. The system is still usable at an error rate of 10^{-2}, but disturbances in the speech signal are clearly present. At an error rate of 3.10^{-2} the speech signal is severely disturbed and the intelligibility is largely lost.

3.4 Residual Excited LP (RELP) Coders

Next to ADPCM, the residual excited linear prediction (RELP) coder is the second kind of speech coder that could be called "modern", in the sense that

- its performance does not depend on pitch and/or voicing detection,
- it uses linear prediction and generates a prediction residual,
- its implementation is completely digital,
- and the transmission signal is a low rate digital bit stream.

Whereas critically (Nyquist) sampled ADPCM requires at least 3 bit per sample for acceptable speech quality, RELP coders can reach good quality at significantly lower bit rates.

Figure 3.15 shows the block diagram of the generic RELP coder. The speech signal $s[n]$ is subject to linear prediction analysis on a segmental basis and it is fed through the inverse filter $A(z)$ to obtain the prediction residual $d[n]$. The low-pass filter LPF only passes the baseband signal so that the sampling frequency can be decimated by a suitable factor D. Assuming a sampling frequency of 8 kHz, the decimation factor is often in the range of 2 to 8. The transmission signal is obtained by merging the appropriately coded baseband (bb) and prediction coefficients (cs).

After decoding these signals in the decoder, the baseband signal is restored to its original sampling frequency by interpolation, the generic form of which consists of up-sampling and appropriate low-pass filtering. Then, the missing high frequencies are regenerated (HFR) and added to the baseband signal to obtain the reconstructed prediction residual $d'[n]$.

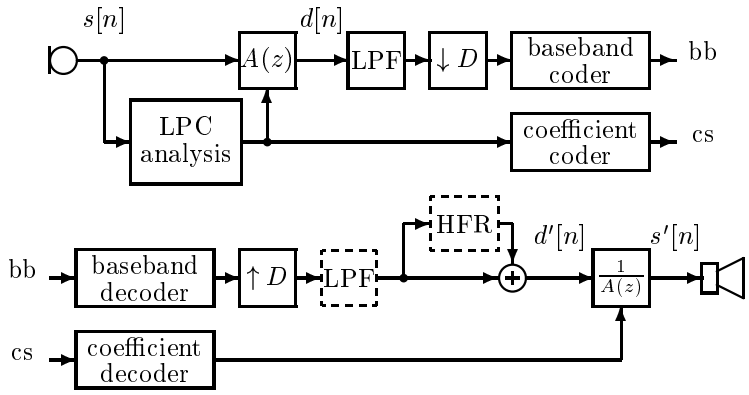

Figure 3.15: Block diagram of the generic RELP encoder (above) and decoder (below).

The HFR function can be realized on the basis of non-linear distortion of the baseband signal as is done in voice excited vocoders. As will be explained in Section 3.4.2, another specific HFR function is obtained by just up-sampling, so that the interpolating low-pass filtering and an explicit HFR function can be omitted, then. Finally, the reconstructed speech signal $s'[n]$ is obtained by exciting a synthesis filter $1/A(z)$ by $d'[n]$. The transfer function of the synthesis filter is regularly updated by the received prediction coefficients.

It is the decimation that brings the greatest reduction in bit rate. Of course, efficient baseband coding exploiting the flat spectral envelope of the prediction residual and noise shaping by the reconstruction filter $1/A(z)$, as discussed in Section 3.2.2, also contributes to bit rate reduction, just as effective coding of the prediction coefficients (recall Section 3.3.3).

The first attempt to encode the prediction residual waveform at a relatively low bit rate was reported by James Dunn of ITT in 1971 [138]. Although this system does not take advantage of a reduced bandwidth baseband, more like a D*PCM coder as already discussed in Section 3.2.2, it is interesting to mention some characteristics of Dunn's approach before we turn to the more relevant application of voice excitation to the prediction residual. Dunn's coder used a second-order adaptive inverse filter which was updated every 10 ms. The sampling frequency was 8 kHz. The prediction residual was coded with forward adaptive sigma-delta modulation (also at 8 kHz) and the total bit rate

was 9600 bit/s. The obtained speech quality was judged to be at least as good as 4-bit logPCM at 32 kbit/s, he wrote.

3.4.1 RELP with non-linear spectral regeneration

The first speech coders that apply voice excitation to the prediction residual are the coder of Un and Magill [139], who introduced the acronym RELP, and the coder indicated by "voice excited predictive coding system" of Atal, Schroeder and Stover [140], both reported in 1975.

The RELP coder of Un and Magill operates at a sampling frequency of 6.8 kHz. The order of the LPC is 10 and the update rate 50 times per second. Each analysis frame is extended 30 samples on each side of the central 20-ms segment, so that subsequent frames have 60 samples overlap. This helps in smoothing the response of the filters $A(z)$ and $1/A(z)$ to updating the prediction coefficients. The analysis frame is multiplied by a Hamming window and 10 reflection coefficients are computed using the autocorrelation method, primarily because it assures stability of the synthesis filter in the decoder. The residual obtained by a direct-form inverse filter is low-pass filtered with a cut-off frequency of 800 Hz and encoded with adaptive delta modulation with hybrid companding at 6.8 kbit/s. The hybrid companding comprises the usual backward adaptation (see Section 3.2.1) as well as forward adaptation on the basis of the prediction residual energy for each frame. This parameter is PCM encoded and conveyed to the decoder at a rate of 200 bit/s, together with a gain factor, encoded in the same way and also at 200 bit/s, to control the output level of the decoded speech signal. The reflection coefficients are PCM encoded using 45 bits per frame. Together with 150 bit/s for frame synchronisation this adds up to a total bit rate of 9600 bit/s. In the decoder the HFR is based on full-wave rectification and subsequent double differencing to enhance the energy of the generated higher frequencies. After high-pass filtering this signal it is added to the baseband together with random noise of at least 5 % of the total excitation energy, to form the reconstructed prediction residual in which the baseband signal itself is preserved. Adding noise improves the speech quality, especially for fricatives. The reported speech quality is "very good" with robustness in the presence of acoustical noise. The hardware complexity is "simpler than any reasonably well-performing pitch extractor in a pitch-excited vocoder", according to Un and Magill.

The RELP coder of Atal, Schroeder and Stover [140] has a bit rate

of only 3600 bit/s due to low bit rate baseband coding. The incoming speech signal which is sampled at 10 kHz for this coder, is low-pass filtered to a 500 Hz baseband. This signal is coded with an ADPCM coder using a forward adaptive 1-bit quantizer and a pitch-only predictor with $a_L z^{-L}$. It should be noted that it is not the baseband of the prediction residual that is sent but the quite narrow baseband directly derived from the input speech signal. The output of the ADPCM coder is simply down-sampled to 1 kHz. The parameter update rate is 50 times per second. The delay parameter L is a 6-bit number and the pitch prediction coefficient $a_L$is quantized to 3 bits. The step size of the 1-bit quantizer is conveyed to the decoder by 3 bits. The rms value of the speech signal is also transmitted using 4 bits. The coefficients of the 12^{th} order linear prediction applied to each 20-ms frame of the input signal are transformed by an inverse hyperbolic tangent function - a log-area like function - then quantized and finally sent with 36 bits for the whole set. Together with 20 bits per frame ADPCM data, the total bit rate adds up to 72 bits/frame. In the decoder the baseband signal is interpolated, using up-sampling to 10 kHz and a 500 Hz low-pass filter, and subsequently decoded in the usual ADPCM way. This signal is subject to non-linear distortion in the form of half-wave rectification, subsequent center clipping and final cleaning by retaining only one peak between two adjacent zero-crossings. The threshold for center clipping is 25% of the peak value of the signal. The reconstructed excitation signal is obtained by spectrally flattening this signal using a separate 12^{th} order adaptive inverse filter. The output speech signal is then obtained in the generic way using this excitation signal and the received and decoded LPC parameters. A speech quality indication of this coder was not given in the paper.

Another approach uses a sub-band coder (SBC), see Section 3.6, for baseband coding. The paper by Esteban, Galand, Mauduit and Menez from 1978 [141] describes a RELP coder with 1 kHz baseband of the prediction residual. The prediction residual is obtained by an 8^{th} order inverse filter operated by the already quantized prediction coefficients, so that the inverse filter in the encoder and the synthesis filter in the decoder are exactly mutually inverse. The SBC uses 6 sub-bands, the outputs of which are quantized on the basis of forward block adaptive PCM, the number of bits per sub-band being allocated adaptively. The bit rate for these bands together could be set to 7200 bit/s. Furthermore, 1800 bit/s were used for the transmission of 8 PARCOR coeffi-

cients and 600 bit/s for an energy parameter indicating the level of the high frequencies to be generated. The HFR is based on rectification and subsequent addition of random noise. Simulation results show that the codec "can maintain a telephone quality without any speaker dependency at transmission rates of 9600 bit/s". Finally the paper reports on an estimation of the required hardware complexity, indicating a processing rate of "about 4 million of elementary 16-bit instructions per second, which can be achieved with available third generation bipolar microprocessors". The use of an SBC for an already spectrally flattened signal is questionable, however, and it might have been better to encode the baseband directly by some form of APCM, for which at 7200 bit/s 3.6 bits/sample are available (See Section 3.6). This is also much simpler than an SBC. This approach does not provide the freedom of allocating different numbers of bits to the various frequency bands, however.

An inaccuracy in high frequency generation by non-linear distortion of the baseband is that the control over the amplitude spectrum of the regenerated high frequencies is limited. The ideal spectral shape is flat, but there is no guarantee at all, that the generally applied non-linearities of rectification and clipping will meet this requirement. In the RELP coder of Atal, Schroeder and Stover this problem was addressed by applying a separate adaptive inverse filter to the regenerated prediction residual. This approach, however, is not optimal in the sense that the regenerated high frequency spectrum does not obey any speech-like all-pole model so that LPC may not optimally flatten the spectrum.

3.4.2 RELP with spectral replication

An important milestone in the development of RELP coders was reached by the formulation of spectral replication techniques for high-frequency regeneration, by Makhoul and Berouti in 1979 [142]. This approach provides optimum control over the regenerated high-frequency amplitude spectrum because the high frequencies consist of replica of the baseband of the optimally flattened prediction residual. Two replication methods were distinguished, i.e. spectral folding and spectral translation. Both methods are starting from the recovered and still decimated baseband signal in the decoder, and the first operation is up-sampling to the sampling frequency of the original speech signal. This is done by inserting $D-1$ zero-valued samples between the baseband samples, where D is the (necessarily integer) decimation factor. This completes the spectral folding method already. The obtained spectrum is depicted in Figure

3.16, for the case $D = 4$. The band from $B = f_s/(2D)$ to f_s/D, where f_s is the original sampling frequency and $0-B$ the baseband, consists of the folded baseband. The whole spectrum consists of the periodic extension of this $0 - f_s/D$ band. The generation of spectrally translated versions of the baseband requires no additional processing for odd multiples of the baseband, as is visible from Figure 3.16. For the translation to even multiples of the baseband, however, additional modulation and filtering is required and in addition, suppression of these bands in the spectrally folded spectrum is necessary before merging the even and the odd bands [142]. Thus, spectral translation is more complex than spectral folding.

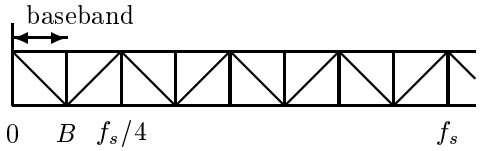

Figure 3.16: Spectral folding scheme for sampling frequency f_s and decimation $D = 4$.

The remaining problem in spectral replication is that for voiced speech the replicated frequencies are generally not harmonics of the pitch. The differences between these frequencies do equal the pitch, however, except at multiples of the frequency B. These harmonic irregularities are audible, more pronounced in case of high pitch as in female speech and less for low pitch voices, as background "hollow, metallic" out-of-tune tones and the phenomenon later became known as "tonal noise". To solve this problem, Makhoul and Berouti suggested a replication which is adaptive to the pitch and thus creates a pure harmonic spectrum. This is most easily done in the frequency domain, according to them, and therefore it might effectively be combined with an appropriate form of transform coding (TC) for the baseband (see Section 3.6).

In 1981 Katterfeldt [143] reports on a RELP coder which can be set to several bit rates, with TC of the 0-800 Hz baseband and a high frequency generation consisting of pitch synchronous spectral translation. The sampling rate of the speech signal is 8 kHz. At 9.6 kbit/s the performance is so good "that 2.5 kbit/s can be spent for additional side information to obtain a better high-frequency regeneration in order

to reduce the remaining HFR-distortions", according to the author. At this bit rate the side information contains the amounts of translation for the several bands to cover the whole spectrum. The amounts of translation are determined on the basis of maximising the normalised cross-correlations between the DFT spectra of the input speech and the translated baseband. At 4.8 kbit/s only the pitch value is transmitted, determined by the cepstrum method.

Note that in this sophisticated approach the pitch is transmitted again, albeit that it is not a critical parameter, anymore. Pitch errors can at most cause translations which have lost the pitch synchronism causing temporarily some background tonal noise.

In 1983 Katterfeldt and Behl present a conference paper [144] in which the performance of a later version of the above coder is discussed. The most important findings are that the coder proved to be robust with regard to acoustical background noise and transmission errors. They wrote: "Simulations with a mobile-radio channel model showed that in case of frequent channel disturbances speech transmitted 'digitally' by RELP in conjunction with a band-limited MSK modulator, has a significantly more acceptable quality than speech which was transmitted by 'analog' FM.". MSK stands for minimum-shift keying [145]. The average SNR on the channel was 14.7 dB, in the digital case resulting in an average bit error rate of 8.10^{-3} at 9.6 kbit/s and half this number at 4.8 kbit/s.

Un and Lee (1982) [146] proposed a RELP coder with a hybrid form of high-frequency regeneration, consisting of extension of the baseband by nonlinear distortion over a fraction of the full-band and then spectral folding of this extended band to recover the full band signal. The rationale behind this approach is that the authors believed that tonal noise is caused by the harmonic irregularities at multiples of the upper edge frequency B of the baseband. By non-linear extension the harmonic structure is essentially maintained. By extension of the 0-800 Hz baseband in this way up to 2000 Hz - using full-wave rectification and double differencing, band-pass filtering to the 800-2000 Hz band and recombination with the baseband - and then applying spectral folding by up-sampling by a factor of two to restore the original sampling frequency of 8 kHz, the number of irregular frequency locations is reduced to just one, thus promising less tonal noise. The price to be paid is additional filtering and no control over the spectral flatness of the nonlinearly generated part. The "combined method yielded better (that is, less hollow

and smoother) synthetic speech quality than either method used alone", according to the authors.

The state of the art in RELP coding in the early 1980s is very well represented by the RELP coder of Viswanathan, Higgens and Russell, published in 1982 [147]. Out of several alternatives they chose their "optimised 9.6 kbit/s baseband coder". This coder operates at a sampling frequency of 6.67 kHz. The incoming speech is analysed using 27-ms segments. Each segment is multiplied by a Hamming window and the autocorrelation method is used to determine 8 reflection coefficients. These are transformed into LARs and quantized using the uniform sensitivity method (see Section 3.3.3). The bit allocation was ordered according to increasing LAR-index 5,5,5,4,4,4,3,3, yielding a total of 33 bits per set. The quantized LARs are transmitted to the decoder but first of all transformed into a-parameters, and used by an inverse filter $A(z)$ to generate the prediction residual from the input speech signal. Subsequently, the residual sequence is low-pass filtered suppressing the band above 1.1 kHz, and down-sampled by a factor of 3. This signal is fed to a baseband coder consisting of an ADPCM coder with a single coefficient pitch-only predictor $a_L z^{-L}$, in which a_L is the coefficient and L the pitch period. The predictor is forward adaptive and a_L and L are transmitted with 4 and 6 bit codes respectively. The quantizer is a (Laplacian Lloyd-Max) nonuniform 3-bit forward adaptive quantizer and the adaptation (gain) parameter is logarithmically quantized and transmitted by a 6-bit code. Together with 30 bits for error protection and 1 bit for frame synchronisation this adds up to a total of 260 bits per 27-ms frame.

In the decoder spectral folding by up-sampling is used and the baseband and the band above it are treated separately. The high band is derived from the randomly perturbed samples of the up-sampled baseband and added to the unperturbed baseband to form the excitation signal for a synthesis filter $1/A(z)$. Perturbation at randomly chosen locations is done by interchanging non-zero samples by adjacent zero-valued samples, but only if the non-zero sample exceeds a certain threshold, which adapts itself slowly to the long-term signal level. This avoids high amplitude pitch related samples to be perturbed.

Perceptually, it was found that the perturbation noise has the effect of masking the tonal noise but it also causes slight roughness in the reconstructed speech. The coder produced good speech quality in the absence of transmission errors and introduced only a slight degradation

in quality for error rates up to 1 percent. Interesting are also some results of a DRT test of a real-time version of this coder: "The DRT scores obtained for high-quality speech input are 89.9 for error-free transmission, 90.8 for 1% channel error, and 83.3 for 5 percent channel error. For error-free transmission the DRT scores are 89.3 using the coder input in a typical office noise environment and 82.2 for an airborne command post noise environment."

Another solution to the tonal noise problem was proposed by Arjmand and Doddington in 1983 [148]. They used a pitch adaptive bandwidth of the baseband such that its bandwidth spans an integer multiple of the pitch frequency. In the decoder, the baseband can then be replicated on the basis of translation, ideally resulting in a pure harmonic spectrum. Although this approach has shown to be able to largely remove the tonal noise, it results in a slightly variable bit rate. At a nominal value of 9600 bit/s, the actual bit rate deviated by a maximum of 0.63% from this value over a time interval of 4.5 seconds of speech.

Still another solution was suggested by Hedelin 1983 [149]. He investigated non-uniform down-sampling of the prediction residual in RELP coding. He compared several alternatives amongst which two schemes which adapted the particular down-sampling scheme to the baseband signal itself by using a mean-square error measure. He concluded, amongst other things, the following: "The two adaptive schemes can be used to significantly improve the high frequency reproduction. The listening tests indicate that a non-uniform scheme with an average decimation of 1:5 roughly corresponds to a uniform 1:4 decimation. These tests were performed with 8 sentences and 5 listeners in A-B comparisons." Although this work indicated that non-uniform down-sampling could bring improved speech quality, it did not get any follow-up.

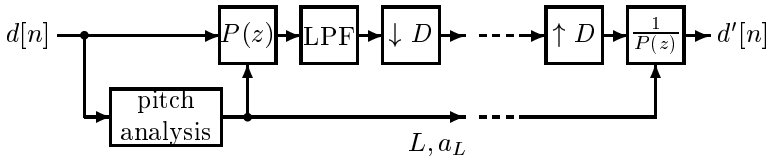

Figure 3.17: Pitch prediction scheme avoiding tonal noise.

Sluijter, Bosscha and Schmitz proposed a solution to the tonal noise problem in their paper "A 9.6 kbit/s Speech Coder for Mobile Radio

Applications" from 1984 [150]. See also the underlying patent [151]. As mentioned before, the tonal noise is caused in periodic (voiced) speech fragments by the inharmonic extension of the harmonic baseband spectrum beyond the baseband, due to the spectral replication process. For non-periodic (unvoiced) speech the spectral replication process causes no unwanted effects. The solution is to feed the prediction residual $d[n]$ first through a pitch predictor $P(z) = 1 - a_L z^{-L}$ by which possible periodicity is removed so that decimation in the encoder and the spectral replication process in the decoder are always operating on non-periodic signals. Figure 3.17 shows such an architecture for the case of spectral folding. After the spectral folding process the full-band residual $d'[n]$ is recovered by the pitch synthesis filter $1/P(z)$. In this way, a correct harmonic spectrum is generated all over the band. Both the values a_L and L are determined from the autocorrelation function $\rho[n]$ of a segment of $d[n]$, the duration of which is for instance 20 ms. The pitch period L is determined by the location of the maximum of $\rho[n]$ in an interval corresponding to 2-10 ms and the prediction coefficient is given by $a_L = \rho[L]/\rho[0]$. For unvoiced speech sounds the correlation peak is small so that a_L becomes small as well. This implies that in these cases the signal passes the pitch predictor almost unaltered. The same situation occurs for pitch periods exceeding 10 ms. For these long pitch periods tonal noise is not perceived anyhow, and exclusion of these long periods from prediction saves considerable computational effort.

Whereas the application of a pitch predictor in a RELP coder was not new, it had exclusively been used to improve the quantization noise performance of the baseband coder. In the new approach it is essential that pitch prediction is applied to the full-band residual prior to the decimation process. As a result, tonal noise is removed and the speech quality is practically enhanced to telephone quality. Another elegant property of this system is that the baseband is kept transparent since

$$A(z) \times P(z) \times 1/P(z) \times 1/A(z) = 1 \,, \tag{3.42}$$

independent of the various parameters. The transparency of the coder makes it very robust with respect to acoustical background noise entering the microphone.

Another issue is the absence of low frequencies below 200 or 300 Hz, for instance due to the microphone. In the spectral folding process this results in spectral gaps at multiples of the decimated sampling frequency. Although this seems to be a disadvantage of spectral folding, it appeared

that the subjective quality of the reproduced speech was not significantly affected (at a sampling frequency of 8 kHz and a decimation factor $D = 4$, in which case such a gap is created at 2 kHz).

3.4.3 Implementations

Some details of an implementation of this coder are discussed in the paper by Vary and Sluijter "Speech Processing in the Mobile Radio Terminal" presented in 1985 [152]. Perhaps the most critical issue in the design is the design of the low-pass filter. It must be a filter with a steep transition between the passband and the stopband to minimize aliasing effects due to the spectral folding process. The design presented in the paper concerned an IIR filter with a 6-dB cutoff frequency at 1000 Hz, for 8 KHz sampling frequency and a decimation factor of 4. It is a sixth-order elliptic (Cauer) filter with a passband ripple of 0.28 dB (see [137], a 0.28 dB cutoff frequency at 940 Hz and a stopband attenuation of >44.7 dB for frequencies exceeding 1155 Hz. The paper reports that the speech quality is not affected by the non-linear phase characteristic as compared to a linear-phase FIR filter with a similar amplitude characteristic. Nothing is reported about time-domain effects, such as ringing. In terms of computational complexity, however, the IIR solution is clearly preferred.

The paper from 1984 [150] also reports on a predecessor of this coder without the pitch predictor, also at 9.6 kbit/s [153], [154], [155] which had been implemented in hardware (see Figure 3.18). The architecture of this implementation is based on an array of signal processors (NEC μPD7720) which can communicate via their 8-bit parallel interfaces over an 8-bit wide data bus. The μPD7720 has a 250 ns instruction-cycle time (4 MIPS), it is packed in a 28-pin house and its power consumption is 1 Watt. Via the serial interfaces of the processors they can communicate with the outside world such as AD and DA converters and modems. To each processor an I/O controller is added to the parallel interface. They communicate in a chain and control the bus traffic. An I/O controller needs only a few tens of logical gates. For the implementation of the speech coder 4 processors are needed in the encoder and 2 in the decoder. In the encoder, the first processor inputs speech samples, does LPC analysis and encodes the resulting coefficients, for even segments. The second processor executes the same tasks for odd segments. The third processor is mainly used for buffering input speech samples for the remaining processing. Finally, the fourth processor executes direct-form

inverse filtering, low-pass filtering, decimation, baseband coding and bit stream generation. In the decoder, the first processor unpacks the bit stream and decodes the baseband signal and the LPC coefficients. The main task of the second processor is the synthesis filtering and output of the synthesised speech samples. The complete codec is mounted on a single 9×4 inch board.

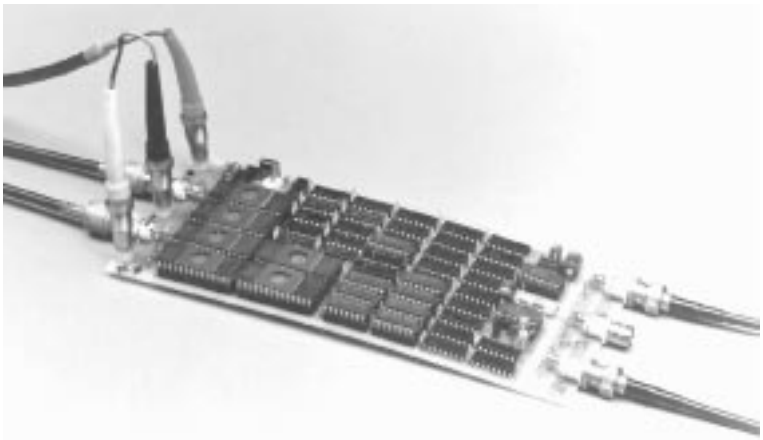

Figure 3.18: RELP-codec board from 1983 with 6 processors μPD7720 (NEC) with UV-erasable (through the round windows) program and data EPROMs. The upright connectors at the left hand side supply power. The other two connectors at the left hand side are the digital output of the encoder to the transmission path and the digital input from the transmission path to the decoder. The connectors at the right hand side are the analog speech input and output of the encoder and decoder, respectively.

This realization has also been used to test the perceptual effect of randomly distributed transmission errors. In the case of bit errors the intelligibility of a speech coder in general will decrease with increasing bit error rate until a certain breaking point beyond which the intelligibility decreases quickly. The breaking point for the coder under test appeared to occur at about 1% error rate. By adding the pitch predictor to the scheme, additional error propagation will result in the decoder. The perceptual effect of this was not investigated separately, but pitch prediction was included in the tests discussed in the next chapter.

Other hardware implementations of RELP coders have been reported. In 1983 [156] a multi-rate speech digitizer implementation based on the

2900 series bit-slice microprocessors of Advanced Micro Devices (AMD) and a TRW multiplier, is described by Lee of the Gold Star Electric Company, Korea and Un, Lee, Shin and Lee of the Korean Advanced Institute of Science and Technology. The system is based on the RELP coder originally proposed by Un and Magill [139], as described previously. The physical size of this versatile system is $19 \times 15 \times 5.25$ inch and its power dissipation amounts to 50 Watts. In 1984 [157], Dankberg, Iltis, Saxton and Wilson of M/A-COM LINKABIT Inc., San Diego, report on a RELP implementation on a single card of 35 square inches using two Texas Instruments TMS320 digital signal processors, one for the encoder and one for the decoder. The power consumption is less than 5 Watts.

3.5 Analysis-by-Synthesis Coders

The application of analysis-by-synthesis techniques to speech signals originates from the 1930s. The particular application in those days was the matching of speech spectra with simple resonance curves. The article of Bell, Fujisaki, Heinz, Stevens and House from 1961 [158], dealing with the reduction of speech spectra by analysis-by-synthesis techniques, shows that it was still a topical subject in those days. In addition the article contains some interesting historical information on the subject.

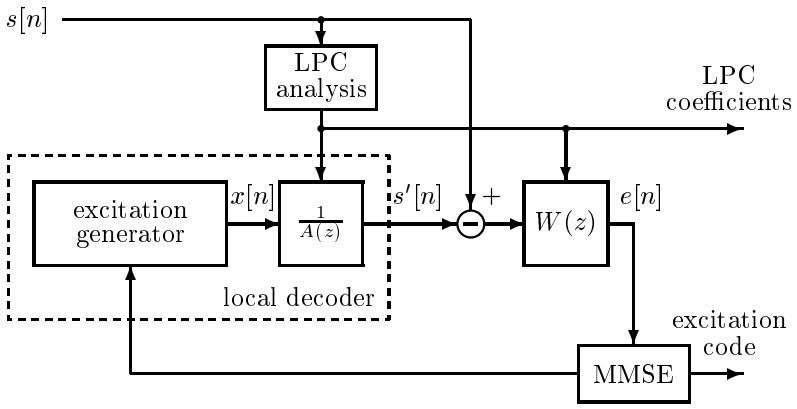

Figure 3.19: Generic analysis-by-synthesis architecture.

The analysis-by-synthesis strategy applied to find bit-rate efficient

substitutes for the prediction residual has been introduced by Schroeder, Atal and Remde in 1981 [159, 160, 161]. Figure 3.19 shows a generic block diagram of the linear prediction based analysis-by-synthesis architecture. A segment of the input speech signal $s[n]$ is subject to LPC analysis resulting in a set of prediction parameters. These are used in a synthesis filter $1/A(z)$ in a local decoder which produces a synthetic speech signal $s'[n]$ with the aid of the excitation signal $x[n]$ from an excitation generator. This generator can generate only a limited number of excitation sequences under control of an excitation code. The synthetic speech segment is subtracted from the original speech segment and the difference is fed through a perceptual weighting filter with transfer function $W(z)$, given by

$$W(z) = \frac{A(z)}{A(z/\gamma)} \; . \tag{3.43}$$

The weighted error $e[n]$ is then minimized in a mean-square error (MMSE) sense by selecting the most suitable excitation sequence. Note that the perceptual weighting also uses the prediction parameters so that the weighting is adaptive to the speech signal. The constant γ is a spectral smoothing factor in the range $0 \leq \gamma \leq 1$, giving some control over the perceptual coding error, see Appendix D. If $\gamma = 1$ then $W(z) = 1$ and there is no weighting giving roughly a spectrally white coding error. If $\gamma = 0$, then $W(z) = A(z)$ and the spectrum of the error in the reconstructed speech signal is proportional to the speech spectrum. In Section 3.2.2 we already mentioned that the perceptual optimum for the spectral shaping of the coding error is somewhere halfway between white and proportional to the speech spectrum itself. The perceptual optimum is quite broad and fine control of γ is not very critical. A typical value of γ is 0.8 at 8 kHz sampling frequency.

The transmitted LPC parameters and excitation code are used in a remote decoder, which is identical to the local decoder, to reconstruct the synthetic speech signal $s'[n]$ there. A low bit rate is achieved by allowing only a limited set of representative excitation sequences so that the excitation code for a segment of speech is of limited size. This still allows a broad class of excitation sequences such as a thinned out sample sequence, not necessarily on a regular basis according to the down-sampling process in RELP coders, but also limited sets of fully populated sequences.

A more elaborate version of the coder uses a pitch synthesiser $1/P(z)$ in front of the synthesis filter $1/A(z)$. Additional analysis for the de-

termination of the pitch predictor parameters will then be part of the encoding process, of course. The pitch parameters will be part of the transmission signal so that they will slightly increase the bit rate.

So, the innovation with respect to RELP coders is that the analysis-by-synthesis framework allows the much broader class of excitation sequences, without the need for one or more baseband low-pass filter(s) and HFR techniques, and perceptual weighting of the coding error.

In [161] work is reported on an elaborated ADPCM coder which contains a center-clipper in front of its quantizer. In this way, several samples of the quantized prediction residual become zero. In one of the experiments the distortion in the reconstructed speech signal was still small if 95% of the quantized residual samples were clipped to zero. This experimental set up can be seen as a direct predecessor of the analysis-by-synthesis multi-pulse excitation method which is the subject of the next subsection.

Another experiment described in the same paper uses a noisy excitation sequence as a substitute for the pitch prediction residual after short-term and long-term inverse filtering of the speech signal. Using a binary tree coding strategy, random-amplitude excitation samples are assigned to subsequent sampling instants. Only the path in the tree that minimises the perceptual mean-square coding error during a segment of the speech signal is retained and transmitted at a cost of 1 bit/sample. It was found that the reconstructed speech signal had "no audible noise". This experiment can be seen as the direct predecessor of analysis-by-synthesis code excitation which will be discussed in Section 3.5.2.

In Section 3.5.3 a third excitation method, analysis-by-synthesis regular-pulse excitation, is described. This was invented by Kroon, Deprettere and Sluijter in 1985 and it is clearly inspired by down-sampling as in RELP coders [162,163]. The difference with down-sampling, however, is that the regular-pulse excitation signal only corresponds to regular down-sampling within a segment of typically 5 ms. The initial phase of the down-sampling process changes from segment to segment, by which tonal noise is prevented.

3.5.1 Multi-pulse excitation (MPE)

In multi-pulse excitation (MPE) coders the excitation signal consists of a number of freely located pulses (samples) with optimised amplitudes in each segment. In the original paper of Atal and Remde [164] from

1982, the weighted error is minimized over segments of 5 or 10 ms. The number of pulses to arrive at RELP-like bit rates is in the order of 10 pulses per 10-ms segment at a sampling frequency of 8 kHz. The rest of the excitation sequence equals zero. The information content of the pulse locations is given by

$$\log_2 \binom{N}{K} = \log_2 \frac{N!}{(N-K)!K!} \text{ bits} , \tag{3.44}$$

where N is the number of possible locations and K the number of MPE-pulses. In the example of 10 pulses per 10 ms and a sampling frequency of 8 kHz this becomes 40.6 bits. Suppose that the amplitudes can be coded at 3 bits per pulse and that a block amplitude parameter of 5 bits is used, then the bit rate of this excitation function is 7.6 kbit/s, when 41 bits are allocated for pulse position coding. Together with the LPC parameters encoded at 40 bits per set and an update rate of once per 20 ms, for instance, the total bit rate becomes 9.6 kbit/s.

The algorithm to determine the optimum pulse positions is a problem, however. An exhaustive search in which $2^{40.6}$ sequences with different pulse positions are evaluated for their perceptual error, is much too complex. If one such evaluation would take 25 cycles on a 250 MHz processor, which were far-future figures in the 1980s, the exhaustive search for one 10-ms segment would take about 2 days. Therefore, a sub-optimal algorithm was proposed by Atal and Remde which searches one pulse position at a time. In this algorithm an initial step processes the memory contents of the filters due to previous excitations, assuming zero excitation in the present segment. This gives rise to an error sequence

$$e_0[n] = e[n]\big|_{x[n]=0} , \quad n = 0, 1, \ldots, N-1 , \tag{3.45}$$

including the contribution of the speech signal $s[n]$ over the current segment. The segment duration is assumed to be N samples. The contribution of a first impulse to the error sequence is evaluated as if there were only one pulse in the excitation signal. Thus, the new error signal is given by

$$e_1[n] = e_0[n] - x[n_1]h[n-n_1] , \tag{3.46}$$

where $x[n_1]$ stands for the amplitude of this first pulse at location n_1 and $h[n]$ for the impulse response of the system $H(z)$ given by

$$H(z) = \frac{1}{A(z)}W(z) = \frac{1}{A(z/\gamma)} . \tag{3.47}$$

The total square error over the segment is now given by

$$E_1 = \sum_{n=0}^{N-1} e_1^2[n] \; , \tag{3.48}$$

which quantity has to be minimized. By substitution of Equation 3.46 into 3.48, determining the partial derivative of E_1 with respect to $x[n_1]$ and setting this to zero, the optimum amplitude

$$x[n_1] = \frac{\sum\limits_{n=0}^{N-1} e_0[n]h[n-n_1]}{\sum\limits_{n=0}^{N-1} h^2[n-n_1]} \tag{3.49}$$

is obtained for the particular location n_1. Trying all N locations and selecting the one with minimum E_1 yields the best solution. If we denote the total square error before the application of the first pulse by E_0 then the new error can be written as

$$E_1 = E_0 - \frac{\left(\sum\limits_{n=0}^{N-1} e_0[n]h[n-n_1]\right)^2}{\sum\limits_{n=0}^{N-1} h^2[n-n_1]} \; . \tag{3.50}$$

Because E_0 and E_1 are both positive quantities, the search could also be based on localising n_1 for the maximum of the last term in this equation. Then, the presence of this pulse with optimised amplitude and location is accounted for in the error sequence according to Equation 3.46 yielding the updated error sequence $e_1[n]$ and a second iteration is carried out as if there were only one pulse in the excitation signal, again. The final solution is then obtained after K iterations. The algorithm is completed by determining the final filter states with the found solutions to prepare the analysis of the next speech segment.

An improvement in this sequential approach consists of re-computation of the amplitudes, using the same pulse locations. The final error is given by

$$E_K = \sum_{n=0}^{N-1} \left(e_0[n] - \sum_{i=1}^{K} x[n_i]h[n-n_i] \right)^2 \; . \tag{3.51}$$

The minimum of E_K is found by setting its partial derivatives with respect to $x[n_k], k = 1, 2, \ldots, K$ to zero. This leads to the set of simultaneous equations

$$\sum_{i=1}^{K} x[n_i]c[n_i, n_k] = v[n_k] \ , \ k = 1, 2, \ldots, K \ , \tag{3.52}$$

where $c[n_i, n_k]$ stands for the covariance coefficients

$$c[n_i, n_k] = \sum_{n=0}^{N-1} h[n - n_i]h[n - n_k] \ , \tag{3.53}$$

and

$$v[n_k] = \sum_{n=0}^{N-1} e_0[n]h[n - n_k] \ , \tag{3.54}$$

and the solution of this system with the known pulse locations yields the re-optimised amplitudes $x[n_k], k = 1, 2, \ldots, K$. The error E_K using these amplitudes can then be written as

$$E_K = E_0 - \sum_{k=1}^{K} x[n_k]v[n_k] \ . \tag{3.55}$$

Thus, the solution is implicitly given by the system $\mathbf{Cx{=}v}$ according to Equation 3.52, where $\mathbf{C}$ is a positive definite $K \times K$ symmetrical covariance matrix because $c[n_i, n_k] = c[n_k, n_i]$ and $\mathbf{x}$ and $\mathbf{v}$ are $K \times 1$ column vectors. This kind of system can efficiently be solved by using the modified Cholesky algorithm, just as in the LPC covariance method (see Section 2.5.1).

In a refined re-computation method for the sequential approach, proposed in [165], re-computation of the pulse amplitudes is already applied after two pulses have been located, and re-computation is repeated after the determination of each next pulse. This makes that during the search for the best location of a $(k+1)^{th}$ pulse, the location of one of the k already allocated pulses can only result in an increased error E_{k+1}. This orthogonality prevents the search from allocating more than one pulse to a particular position.

Also in [165], the application of single-tap pitch prediction is investigated. The pitch synthesis filter according to

$$\frac{1}{P(z)} = \frac{1}{1 - a_L z^{-L}} \tag{3.56}$$

is inserted, in the diagram of Figure 3.19, in front of the synthesis filter $1/A(z)$. The search algorithm does not change very much if the system impulse response $h[n]$ of the filter $1/A(z/\gamma)$ is replaced by the impulse response of the cascade of the filters $1/P(z)$ and $1/A(z/\gamma)$. The pitch prediction parameters a_L and L can be found in the way discussed in Section 2.5.5, from the (separately generated) short-term prediction residual of the input speech signal. Alternatively, the pitch period L can be determined with one of the pitch detection methods discussed in Section 3.3.4. This approach is often referred to as open-loop estimation of the pitch parameters. In the analysis-by-synthesis framework, it is also possible to determine these parameters in a closed-loop form. This can be done in the initial step of the processing of a segment while determining $e_0[n]$. While there is zero excitation from the excitation generator, the optimal value of a_L is computed to minimize E_0 for every possible value of L, in a similar way as for the excitation pulses. The value of L giving minimum E_0 results in the optimal values of the pitch prediction parameters. A usual minimum value of the pitch period is 2 ms but the possible values of L are limited to $L \geq N$, however, because recursions in the pitch synthesis filter lead to severe complications in the analysis. This means that for pitch periods shorter than the segment length, L will span multiple periods. This may be a good reason to prefer 5-ms segments instead of 10 ms. The advantage of closed-loop analysis, however, is the better performance because it is better tuned to the analysis-by-synthesis framework including the fact that the pitch parameters are determined using the perceptually-weighted error measure. In [165] it was found that without pitch prediction about 8 pulses per pitch period were needed for high quality speech synthesis. This implies a higher bit rate for high pitch voices such as female voices as compared to male voices. The SNR of the resynthesised speech with regard to the original speech differed more than 6 dB for female and male speech at the same pulse rate. By the application of closed-loop pitch prediction the SNR for female voices could be improved by more than 2 dB. A similar figure is also found by Kroon and Deprettere [166] using open-loop pitch analysis. They also emphasise the significant *perceptual* quality improvement obtained by pitch prediction.

3.5.2 Code excited LP (CELP)

In 1984 Atal and Schroeder reported on an experiment with stochastic code excitation [167]. They used an excitation generator as shown

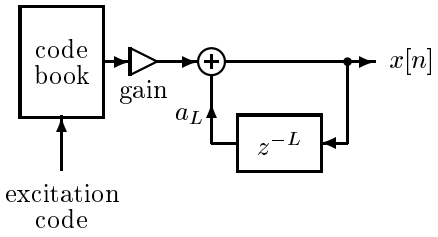

Figure 3.20: Excitation generator of CELP.

in Figure 3.20. The 10-bit excitation code selects one of the 1024 sequences which are stored in the codebook. Each sequence is 40 samples long, which corresponds to a duration of 5 ms at 8 kHz sampling frequency. So, the bit rate for the excitation code is 1/4 bit per sample or 2 kbit/s. The stored sequences are taken from a zero-mean unit-variance white Gaussian random process. They represent the shapes of the waveforms. The selected signal is scaled by a gain factor and then fed to a pitch synthesis filter to obtain the excitation sequence $x[n]$. This architecture has been motivated by earlier work on stochastic tree coding [161] which showed that the prediction residual after two stages of prediction, i.e. short-term and subsequent long-term prediction, had noise-like properties. Its long-term power spectrum is approximately white and the probability density function of the prediction residual sample amplitudes is close to a Gaussian distribution.

Each sequence is scaled to match the speech signal in the analysis-by-synthesis framework, and the perceptual error with respect to the original speech over the 5-ms segment is determined. In an exhaustive search the sequence giving the minimum perceptually weighted error is selected.

It was found from informal listening tests that the synthetic speech matched the original speech very well and that only small differences were noticeable between the high quality synthetic speech and the original speech. The coding procedure took 125 seconds of Cray-1 CPU time to process 1 second of speech. The Cray-1 was the fastest computer available in 1984. So, real-time implementation on the much slower low-power signal processors of the time was completely out of the question. Nevertheless, a lower bit rate as compared to an MPE coder was promised.

In [168], the title of which mentions for the first time the acronym CELP, it is mentioned that the optimum gain factor for each candidate sequence is determined to minimize the perceptual coding error. This can be further specified if we denote the contribution to the error sequence $e[n]$ of a codebook sequence $x_i[n]$, $i = 1, 2, \ldots, I$ for unity gain by $y_i[n]$, $i = 1, 2, \ldots, I$, assuming that there are I codebook sequences in total. If, in addition, we denote the gain factor by g_i then the error can be written as

$$E_i = \sum_{n=0}^{N-1} \left(e_0[n] - g_i y_i[n] \right)^2 \ , \tag{3.57}$$

where $e_0[n]$ is defined by Equation 3.45. By setting the partial derivative of E_i with respect to g_i to zero, its optimum value is obtained as

$$g_i = \frac{\sum\limits_{n=0}^{N-1} e_0[n] y_i[n]}{\sum\limits_{n=0}^{N-1} y_i^2[n]} \ . \tag{3.58}$$

By substitution of this optimum value into Equation 3.57 we find that

$$E_i = E_0 - \frac{\left(\sum\limits_{n=0}^{N-1} e_0[n] y_i[n] \right)^2}{\sum\limits_{n=0}^{N-1} y_i^2[n]} \ . \tag{3.59}$$

In the search for minimum E_i, $i = 1, 2, \ldots, I$, the maximum of the last term in this equation can be sought, equivalently.

3.5.3 Regular-pulse excitation (RPE)

The third excitation method announced in the heading of this section on analysis-by-synthesis coders is regular-pulse excitation (RPE). In RPE [163] the excitation has the form of a down-sampled sequence in a segment of, say again, N samples with down-sampling factor D, according to

$$n_k = \phi + (k-1)D \ , \ k = 1, 2, \ldots, K \ , \tag{3.60}$$

where n_k are the pulse positions for K pulses and $0 \leq \phi \leq D - 1$ is the initial phase, or grid position, of the down-sampling process. Typical values of D are 3 and 4. See Figure 3.21. Usually, $K = N/D$, which

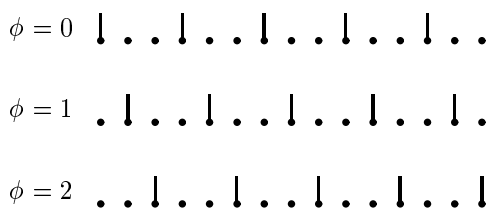

Figure 3.21: RPE pulse-position grids for $D = 3$ and $N = 15$.

requires N to be an integer multiple of D. This is not a rigid constraint and there are ways out if this requirement cannot be met, as will become evident in the next chapter. For each segment of typically 5 ms, all D grids are considered in the error minimization process and the amplitudes of the pulses on each grid are determined so as to minimize the perceptual error. The grid position giving the minimum perceptual error is then selected. For the first time we now have a coding system that can be applied in an optimum way, without the need to fall back on suboptimum solutions because of unmanageable complexity. As compared to MPE, less bits are spent on pulse-position coding, only $\log_2 D$ bits, and consequently more pulses can be used for the same bit rate, but with less freedom in pulse locations. As compared to RELP with a down-sampling factor D, tonal noise is now prevented by the implicitly changing grid positions in subsequent 5-ms segments. If the duration of the segments is chosen too long, tonal noise may be reintroduced. The issue of the design of the low-pass filter in RELP does not exist anymore thanks to the analysis-by-synthesis approach.

The pulse amplitudes $x[n_k]$ at the corresponding locations n_k are given by Equations 3.52, 3.53 and 3.54, but the pulse positions n_k are now regularly spaced. This has major consequences, as will become evident in the next chapter. The error for each grid position is evaluated according to Equation 3.55 and the final grid position is selected on the basis of the minimum of these errors.

RPE has been extensively described in the Ph.D. thesis of Kroon including the combination with pitch prediction [169]. Adoul [170] independently proposed a similar scheme but he used a suboptimal grid positioning. His grid position was not determined by minimum percep-

tual error, but by the position of the first MPE pulse in the sequential approach.

3.6 Other coding paradigms

During the 1970s other coding paradigms have been investigated. Most of them are characterised by their operation in the frequency domain. In this section we will briefly outline the generic forms of sub-band and transform coders on the one hand and the parametric phase vocoder on the other hand. These frequency-domain coders have turned out to be less effective for narrow-band low bit rate speech coding. Another, just emerging, coding technique was vector quantization. The significance of this technique for practical speech coding became evident just after the mid 1980s, as discussed in Chapter 5.

3.6.1 Sub-band and Transform Coders

One of the earliest transform coders applied to speech signals (they were first applied to video signals) was reported by Schafer and Rabiner in 1973 [171]. It was based on the discrete-time Fourier transform. In a transform coder the input signal is split up in equally long, usually overlapping, segments. From each (windowed) segment the actual frequency transform is determined resulting in a short-time spectrum. The frequency-domain signals resulting from the sequence of successive transforms can be interpreted as the outputs of a filter-bank which have been sub-sampled to a sampling rate equal to the rate of the successive transforms. The individual filter characteristics are determined by the applied window shape. The frequency-channel signals are subject to quantization and transmission. Decoding is achieved by the inverse transforms of the short-time spectra and applying overlap-add operations to reconstruct the original time signal. A characteristic property of the transform is that it has a relatively high resolution in the frequency domain with typically 32-256 channels, giving fine control over the frequency distribution of the coding error in formant regions and possibly even at pitch-harmonic locations. Between several adjacent channels, however, significant overlap exists.

In a sub-band coder, on the other hand, the number of frequency channels is relatively low with a (sometimes very) small overlap. The first sub-band coder was proposed by Crochiere, Webber and Flanagan

in 1976 [172]. This proposal concerned 4 sharp-edge sub-bands to cover the frequency range from 200 to 3200 Hz.

Issues in frequency-domain coding are critical sampling, i.e. the number of samples per time unit in the frequency domain to be transmitted equals the number of samples per time unit in the original signal, preferably without aliasing effects, and the adaptive allocation of bits to the frequency samples to establish a substantial bit rate reduction while retaining a required level of perceptual speech quality. The state of the art in 1979 is described in [173], [174]. Filter-bank techniques were well developed in 1983 [175] and the state of the art in 1984 is well reflected in the book of Jayant and Noll [15].

As in all coders, there are two sources of benefit aiming at a low bit rate. First, there is the redundancy in the signal, also often interpreted as the predictability of samples. This is typically represented in the time domain by correlation between samples and in the short-time spectrum by a non-flat envelope. Removal of redundancy gives rise to a lower bit rate. Second, there is the irrelevancy in certain details of the information to be conveyed. Often, this is exploited by quantization which introduces, ideally irrelevant, distortions. Another example of irrelevancy is the much more effective thinning out of the prediction residual, as discussed in the previous sections. An important aspect of the irrelevancy in speech coding is its perceptual nature. Both aspects, redundancy and perceptual irrelevancy in frequency-domain coding will now be evaluated and compared to time-domain coders.

A measure of spectral flatness which is instrumental in the evaluation of redundancy (removal) is the average log normalised spectrum [176] according to

$$\mathrm{spf}(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left\{ \frac{|X(e^{j\theta})|^2}{\rho_x[0]} \right\} d\theta \,, \tag{3.61}$$

in which $\mathrm{spf}(x)$ stands for the spectral flatness of the signal $x[n]$, $\rho_x[0]$ for its zero-lag autocorrelation function representing the energy in the signal, and $|X(e^{j\theta})|^2$ for its energy density spectrum. The latter three expressions are related by Parseval's theorem according to

$$\rho_x[0] = \sum_{n=-\infty}^{\infty} x^2[n] = \frac{1}{2\pi} \int_{-\pi}^{\pi} |X(e^{j\theta})|^2 d\theta \,. \tag{3.62}$$

All these quantities become finite if $x[n]$ is a segment of speech of finite duration. The range of $\mathrm{spf}(x)$ has a one-sided limit 0 for a perfectly

flat spectrum $|X(e^{j\theta})|$ =constant and spf$(x) < 0$ for non-flat spectra. A more convenient form $\mathcal{F}_x$ with a closed interval $0 \leq \mathcal{F}_x \leq 1$ is obtained by

$$\mathcal{F}_x = \exp[\text{spf}(x)] = \frac{\exp\left[\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|X(e^{j\theta})|^2 d\theta\right]}{\frac{1}{2\pi}\int_{-\pi}^{\pi} |X(e^{j\theta})|^2 d\theta}\,, \tag{3.63}$$

with $\mathcal{F}_x = 1$ for a perfectly flat spectrum. For the applicability of this spectral flatness measure to the transform coder a discrete approximation to $\mathcal{F}_x$ can easily be derived [15, 176]. To this end we divide the frequency band $0 \leq \theta \leq \pi$ into K equal bands and we approximate the energy density spectrum by samples located at the centers of these bands and having the values $|X_k|^2$, $k = 1, 2, \ldots, K$, representing the energies in these bands. By using these discrete values and replacing the integrals in Equation 3.63 by summations, we find that

$$\mathcal{F}_x \simeq \frac{\exp\left[\frac{1}{K}\sum_{k=1}^{K} \ln|X_k|^2\right]}{\frac{1}{K}\sum_{k=1}^{K} |X_k|^2} = \frac{\left[\prod_{k=1}^{K} |X_k|^2\right]^{\frac{1}{K}}}{\frac{1}{K}\sum_{k=1}^{K} |X_k|^2}\,. \tag{3.64}$$

This approximation to the spectral flatness apparently consists of the ratio of the geometric and the arithmetic means of $|X_k|^2$, and the approximation improves for increasing K.

By applying the spectral flatness measure to the linear prediction case we find a remarkable result. Consider the stable LP inverse filter $A(z)$ with input $X(z)$ and output $E(z)$:

$$E(z) = A(z)X(z)\,. \tag{3.65}$$

Because the average value of the log spectrum of $A(e^{j\theta})$ equals zero [176] we can write

$$\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|E(e^{j\theta})|^2 d\theta = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln|X(e^{j\theta})|^2 d\theta\,. \tag{3.66}$$

Application of Equation 3.63 to this equality yields the result that the ratio of the spectral flatness measures of the output and input signals equals the inverse ratio of their respective energies

$$\frac{\mathcal{F}_e}{\mathcal{F}_x} = \frac{\rho_x[0]}{\rho_e[0]} = G_P\,. \tag{3.67}$$

Because of Equation 3.17 this energy ratio also equals the prediction gain G_P. So, we can draw the important conclusion that the increased spectral flatness in frequency-domain coding by putting appropriate weightings to the frequency channels gives the same redundancy removal as linear prediction inverse filtering giving the same increase in spectral flatness. Consequently, the maximum coding gain of a frequency-domain coder as well as the maximum prediction gain in a time-domain predictive coder both equal $1/\mathcal{F}_x$. Equation 3.67 also shows that minimization of $\rho_e[0]$ maximises $\mathcal{F}_e$, as announced in Section 2.5.2.

Perceptual irrelevancy in narrow-band frequency-domain coding has been sought in the application of unequal bandwidth channels like the critical bands in human auditory perception, in adaptive bit allocation strategies dedicated to each band and in the application of vector quantization techniques. However, a method which could outperform the bit-rate efficiency and perceptual quality of thinning out the residual in linear predictive coding, has never turned up. That is why frequency domain coders have not become first choice in narrow-band low bit rate speech coding.

3.6.2 Phase Vocoder

The phase vocoder was proposed by Flanagan and Golden in 1966 [177]. In this vocoder, sub-band signals are analysed and represented by time-varying amplitude and phase-derivative spectra. Each sub-band signal $s_k(t)$ is resynthesised according to

$$s_k(t) = A_k(t) \cos\left(\omega_k t + \int_0^t \dot{\varphi}_k(t)dt\right) , \tag{3.68}$$

where A stands for amplitude, ω_k for the central frequency of sub-band k and $\dot{\varphi}$ for the phase derivative (compare with Section 2.3.3). A digital implementation has been reported in 1976 by Portnoff [178]. The ratio of speech quality to bit rate never reached the level of competitive coding paradigms but the approach has been used to perform frequency-scale modifications. In combination with sample rate conversion this can also yield time scale modifications, speeding up as well as slowing down, without altering the pitch and spectral characteristics of the voice.

3.7 Summary

It has been described that in 1937 Reeves' invention of PCM, consisting of the triple process of sampling, quantization and binary coding, started the new era of digital signal processing. In the beginning of this era, the processing exclusively concerned narrow-band speech and music signals. Six to eight bit per sample at a rate of 8000 samples per second were required to provide sufficient perceptual quality. At the same time the channel vocoder, capable of resynthesising speech from a small set of relatively slowly varying parameters, was invented by Dudley. Soon, these techniques were combined and in the mid forties a vocoder at 1.6 kbit/s, including non-uniform logarithmic quantization and cryptography, was in practical use on transatlantic short-wave radio for the purpose of military secure speech communication! The speech quality was poor, however. In those days also the formant vocoder, promising even lower bit rates, was proposed.

In the late forties, delta modulation was invented and the alternative form of differential coding, differential PCM (DPCM), was invented in 1952. All quantization variants with forward and backward adaptive step-size control, both at a sample-by-sample as well as at a syllabic rate, and for DPCM also with forward and backward adaptive prediction, have been discussed. The discussion also included the open-loop variant D*PCM and the noise-feedback coder (NFC) variant. It has been noted that the application of adaptive quantization to PCM started not until the late sixties.

Whereas delta modulation was applied occasionally, like in the space shuttle program at the bit rates of 24 and 32 kbit/s in the mid seventies, PCM at 64 kbit/s became the ITU standard for digital switching and transmission of speech signals in telephony in 1972. The ITU standardized adaptive DPCM (ADPCM) at 32 kbit/s in 1984.

In the meantime, the successful rise of vocoders based on linear-predictive coding (LPC) pushed aside the tedious development of channel and formant vocoders. In 1982, LPC vocoders at 2.4 kbit/s had been developed with reliable digital pitch and voicing detection techniques giving fair (communication) speech quality, at least on clean input speech signals. LPC vocoders and especially the quantization of the LPC-coefficients have been discussed and several pitch and voicing detection techniques have been described and evaluated, also on the basis of the rediscovered "theorem of Sondhi". The main application area still is in military secure speech communications.

It has been described how in the much more robust voice-excited vocoders pitch and voicing detection is avoided by transmitting a relatively narrow baseband in its natural form. The bit rates of this kind of vocoder varied in the range of 4.8–9.6 kbit/s. The quality limiting problems of non-linear generation of the full-band excitation signal from that baseband have been pointed out. The linear-prediction based successor of the voice-excited vocoder, the residual-excited linear predictive (RELP) coder, solved most quality problems by using spectral folding and pitch prediction and the complexity of RELPs turned out to be manageable as well. The major problem left was to seize upon the elusive design criteria for the baseband low-pass filter.

Then, the class of analysis-by-synthesis coders which arose in the early eighties has been described and the multi-pulse excitation (MPE), regular-pulse excitation (RPE) and code-excited linear prediction (CELP) variants were introduced. A detailed analysis of these methods is presented in the next chapter resulting in a so called explicit-form RPE solution with a RELP-like structure and complexity, but with an inherent solution to the low-pass filter design problem. The bit rates are around 10 kbit/s.

Finally, we concluded that the time-domain speech coding paradigm based on linear prediction outperforms frequency-domain coding paradigms such as sub-band coding, transform coding and phase vocoding.

References

1. E. M. Deloraine and A.H. Reeves, The 25th anniversary of pulse code modulation, *IEEE Spectrum*, May 1965 pp. 56–63.
2. A.G. Bell, Improvement in Telegraphy, *USA patent*, No. 174 465, 1876.
3. R. Bernzen, *Das Telephon von Philipp Reis*, LIBRI, ISBN 3-00-004284-9, Marburg 1999.
4. A.H. Reeves, Telecommunications of the future with pulse code modulation, *Canadian Electronics Engineering*, December 1968, pp. 54–56.
5. Systèmes de signalisation électriques, *French patent*, No. 852.183, filed 3 October 1938, granted 23 October 1939, published 5 Januari 1940.
6. A.H. Reeves, Electric Signalling System, *USA patent*, 2 272 070, 1942.
7. W.M. Goodall, Telephony By Pulse Code Modulation, *Bell System Technical Journal*, Volume 26 No.3, July 1947, pp. 395–409.
8. R.C. Mathes and S.B. Wright, The "Compandor" - An Aid Against Radio Static, *Transaction of the American Institute of Electrical Engineers*, June 1934, pp. 860–866.
9. L.A. Meacham and E. Peterson, An Experimental Multichannel Pulse Code Modulation System of Toll Quality, *Bell System Technical Journal*, Volume 27, No.1, Januari 1948, pp. 1–43.
10. B.M. Oliver, J.R. Pierce and C.E. Shannon, The Philosophy of PCM, *Proceedings of the IRE*, Volume 36, November 1948, pp. 1324–1331.
11. S.P. Lloyd, Least Squares Quantization in PCM, *IEEE Transactions on Information Theory*, Volume IT-28, No.2, March 1982, pp. 129–137.
12. J. Max, Quantizing for Minimum Distortion, *IRE Transactions on Information Theory*, Volume IT-6, March 1960, pp. 7–12.
13. *ITU Recommendation* G.711, 1972.
14. K.W. Cattermole, *Principles of pulse code modulation*, Iliffe Books Ltd., London, 1969.
15. N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1984.

16. N.S. Jayant, Adaptive Quantization With a One-Word Memory, *Bell System Technical Journal*, Volume 52, No.7, September 1973, pp. 1119–1144.

17. L.S. Golding, Study of an Adaptive Quantizer, *Proceedings of the IEEE*, Volume 55, No.3, March, 1967, pp. 293–297.

18. R.M. Wilkinson, An Adaptive Pulse Code Modulator for Speech, *IEEE International Conference on Communications*, 1971, Montreal, Canada, pp. (1-11)–(1-15).

19. M.G. Croll, M.E.B. Moffat and D.W. Osborne, 'Nearly Instantaneous' Digital Compander for Transmitting Six Sound-Programme Signals in a 2.048 Mbit/s Multiplex, *Electronics Letters*, 12th July, Volume 9, No.14, 1973, pp. 298–300.

20. A. Croisier, Progress in PCM and Delta Modulation: Block-Companded Coding of Speech Signals, *Zürich Seminar*, 1974, pp. B1(1)–B1(4).

21. M. Deloraine, S. van Mierlo and B. Derjavitch, Méthode et système de transmission par impulsions, *French Patent* No. 932.140, filed 10 August 1946, granted 17 November 1947, published 12 March 1948.

22. M. Deloraine, S. van Mierlo and B. Derjavitch, Communication system utilizing constant amplitude pulses of opposite polarities, *USA Patent*, No. 2 629 857, 1953, filed 8 October, 1947.

23. J.F. Schouten, F. de Jager and J.A. Greefkes, Système de modulation d'impulsions ainsi qu'émetteurs et récepteurs utilisables dans ce système, *French Patent* No. 987.238, filed 23 may 1949, granted 11 April 1951, published 10 August 1951, first filed at 22 May 1948 in the Netherlands.

24. J.F. Schouten, F. de Jager and J.A. Greefkes, Pulse modulation system for transmitting the change in the applied wave-form, *USA Patent*, No. 2 662 118, 1953, first filed at 22 May 1948 in the Netherlands.

25. J.F. Schouten, F. de Jager and J.A. Greefkes, Delta modulation, a new modulation system for telecommunication, *Philips Technical Review*, Volume 13, No.9, March 1952, pp. 237–245.

26. F. de Jager, Deltamodulation, a method of P.C.M. transmission using the 1-unit code, *Philips Research Reports*, Volume 7 Nr.6, December 1952, pp. 442–466.

27. H. Inose and J. Murakami, A Telemetering System by Code Modulation - $\Delta - \Sigma$ Modulation, *IRE Transactions on Space Electronics and Telemetering*, Volume SET-8, September 1962, pp. 204–209.

28. P.J.A. Naus, E.C. Dijkmans, E.F. Stikvoort, A.J. McKnight, D.J. Holland and W. Bradinal, A CMOS Stereo 16-bit D/A Converter for Digital Audio, *IEEE Journal of Solid-State Circuits*, Volume SC-22, No.3, June 1987, pp. 390–395.

29. M.R. Winkler, High Information Delta Modulation, *IEEE International Convention Record*, 1963, pp. 260–265.

30. J.A. Greefkes, Delta Modulation Signal Transmission System, *USA Patent*, 3 249 870, 1966, filed for the first time in the Netherlands at 20 July 1961.

31. J.A. Greefkes and F. de Jager, Continuous delta modulation, *Philips Research Reports*, Volume 23, April 1968, pp. 233–246.

32. H.R. Schindler, Delta Modulation, *IEEE Spectrum*, Volume 7, October 1970, pp. 69–78.

33. J.A. Greefkes and K. Riemens, Code modulation with digitally controlled companding for speech transmission, *Philips Technical Review*, Volume 31, No.11/12, 1970, pp. 335–353.

34. R.L. Auger, M.W. Glancy, M.M. Goutmann and A.L. Kirsch, The Space Shuttle Ground Terminal Delta Modulation System, *IEEE Transactions on Telecommunications*, Volume COM-26, No.11, November 1978, pp. 1660–1670.

35. J.W. Glasbergen, Second generation DELTAMUX, *Philips Telecommunication Review*, Volume 43, No.3, September 1985, pp. 193–201.

36. V.C. Welch, T.E. Tremain, J.P. Campbell, A Comparison of U.S. Government Standard Voice Coders, *Proceedings of the IEEE MILCOM*, 1989, Paper 13.1, pp. 269–273.

37. C.K. Un and H.S. Lee, A Study of the Comparative Performance of Adaptive Delta Modulation Systems, *IEEE Transactions on Communications*, Volume COM-28, No.1, January 1980, pp. 96–101.

38. C.C. Cutler, Differential Quantization of Communication Signals, *USA Patent*, No. 2 605 361, 1952, filed at 29 June 1950.

39. B.S. Atal and M.R. Schroeder, Adaptive Predictive Coding of Speech Signals, *Bell System Technical Journal*, October 1970, pp. 1973–1986.

40. M.D. Paez and T.H. Glisson, Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems, *IEEE Transactions on Communications*, Volume COM-20, April 1972, pp. 225–230.
41. J.D. Gibson, S.K. Jones and J.L. Melsa, Sequentially Adaptive Prediction and Coding of Speech Signals, *IEEE Transactions on Communications*, Volume COM-22, No.11, November 1974, pp. 1789–1797.
42. C.C. Cutler, Quantized Transmission with Variable Quanta, *USA Patent*, No. 2 724 740, 1955, filed at 29 June 1950.
43. P. Cummiskey, N.S. Jayant and J.L. Flanagan, Adaptive Quantization in Differential PCM Coding of Speech, *Bell System Technical Journal*, September 1973, pp. 1105–1118.
44. N.S. Jayant, Digital Coding of Speech Waveforms: PCM, DPCM, and DM Quantizers, *Proceedings of the IEEE*, Volume 62, May 1974, pp. 611–632.
45. W.R. Daumer, P. Mermelstein, X. Maitre and I. Tokizawa, Overview of the ADPCM Coding Algorithm, *IEEE Global Telecommunications Conference (Globecom)*, 1984, pp. 23.1.1–23.1.4.
46. B.M. Oliver, Efficient Coding, *Bell System Technical Journal*, July 1952, pp. 724–750.
47. J.G. Dunn, An Experimental 9600-bits/s Voice Digitizer Employing Adaptive Prediction, *IEEE Transactions on Communication Technology*, Volume COM-19, No.6, December 1971, pp. 1021–1032.
48. J.L. Flanagan, *Speech Analysis Synthesis and Perception*, Springer-Verlag, Berlin, Second Edition, 1972.
49. B.S. Atal and M.R. Schroeder, Predictive Coding of Speech Signals and Subjective Error Criteria, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No.3, June 1979, pp. 247–254.
50. P. Noll, On Predictive Quantizing Schemes, *Bell System Technical Journal*, Volume 57, No.5, May-June 1978, pp. 1499–1532.
51. A. B. Carlson, *Communication Systems*, McGraw-Hill Book Company, 1986.
52. C.C. Cutler, Transmission System Employing Quantization, *USA Patent*, No. 2 927 962, 1960, filed at 26 April 1954.
53. E.G. Kimme and F.F. Kuo, Synthesis of Optimal Filters for a Feedback Quantization System, *IEEE Transactions on Circuit Theory*, September 1963, pp. 405–413.

54. H.W. Dudley, Signal Transmission, *USA Patent*, No. 2 151 091, 1939, filed at 30 October 1935.
55. H. Dudley, Synthesizing Speech, *Bell Laboratories Record*, Volume 15, No.4, December 1936, pp. 98–102.
56. H. Dudley, The Vocoder, *Bell Laboratories Record*, Volume 18, No.4, December 1939, pp. 122–127.
57. M.D. Fagen, Ed., *A History of Engineering and Science in the Bell System*, National Service in War and Peace (1925-1975), Bell Telephone Laboratories Inc., 1978, Chapter IV: Secure Speech Transmission, pp. 291–317.
58. J.V. Boone and R.R. Peterson, *The Start of the Digital Revolution: SIGSALY Secure Digital Voice Communications in World War II*, http://www.nsa.gov/wwii/papers/start_of_digital_revolution.htm, June 2003.
59. W.R. Bennett, Secret Telephony as a Historical Example of Spread-Spectrum Communication, *IEEE Transactions on Communications*, Volume COM-31, No.1, January 1983, pp. 98–104.
60. J.P Campbell Jr. and R.A. Dean, A History of Secure Voice Coding, *IEEE Signal Processing Magazine*, May 1998, pp. 31–32.
61. R.J. Halsey, J. Swaffield, Analysis-Synthesis Telephony, with Special Reference to the Vocoder, *The Journal of Institution of Electrical Engineers*, Volume 95, Part III, 1948, pp. 391–411.
62. F. Vilbig and K.H. Haase, Some Systems for Speech-Band Compression, *The Journal of the Acoustical Society of America*, Volume 28, No.4, July 1956, pp. 573–577.
63. B. Gold and C.M. Rader, The Channel Vocoder, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-15, No.4, December 1967, pp. 148–161.
64. T. Bially and W.M. Anderson, A Digital Channel Vocoder, *IEEE Transactions on Communication Technology*, Volume COM-18, No.4, August 1970, pp. 435–442.
65. D.P. Fulghum, Output Spectrum Contour Scaling for an All Digital Channel Vocoder, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, 1976, pp. 99–102.
66. H.J. Kotmans, A. van Leeuwaarden and R.J. Sluijter, 2400 bits/sec. Channel Vocoder with Harmonic Sieve Pitch Extractor, *Philips Research Laboratories Internal Report*, TN 204/79, September 1980.
67. J.N. Holmes, The JSRU channel vocoder, *IEE Proceedings*, Volume 27, Part F, No.1, February 1980, pp. 53–60.

68. G.J. Bosscha and R.J. Sluijter, DFT-Vocoder Using Harmonic-Sieve Pitch Extraction, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 1952–1955.
69. N.G. Kingsbury and W.A. Amos, A Robust Channel Vocoder for Adverse Environments, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Denver, 1980, pp. 19–22.
70. J.A. Feldman, A Compact Digital Channel Vocoder Using Commercial Devices, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 1960–1963.
71. G.J. Bosscha, R.J. Sluijter, J.M.P.T. Schmitz and H.J. Kotmans, A DFT Vocoder for Transmission of Speech at 2400 bit/s, *Philips Research Laboratories Internal Report*, No. 5963, August 1984.
72. J.P. Parker, Achieving narrow-band secure voice, *Communication International*, August 1986, pp. 41–44.
73. W.A. Munson and H.C. Montgomery, A Speech Analyzer and Synthesizer, *The Journal of the Acoustical Society of America*, Volume 22, No.5, September 1950, p. 678 (Abstract).
74. J.L. Flanagan and A.S. House, Development and Testing of a Formant-Coding Speech Compression System, *The Journal of the Acoustical Society of America*, Volume 28, No.6, November 1956, pp. 1099–1106.
75. B.P. Bogert, On the Band Width of Vowel Formants, *The Journal of the Acoustical Society of America*, Volume 25, No.4, July 1953, pp. 791–792.
76. H.K. Dunn, Methods of Measuring Vowel Formant Bandwidths, *The Journal of the Acoustical Society of America*, Volume 33, No.12, December 1961, pp. 1737–1746.
77. J.L. Flanagan, Automatic Extraction of Formant Frequencies from Continuous Speech, *The Journal of the Acoustical Society of America*, Volume 28, No.1, January 1956, pp. 110–118.
78. J.D. Markel, Digital Inverse Filtering - A New Tool for Formant Trajectory Estimation, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-20, 1972, pp. 129–137.
79. C.K. Un, A Low-Rate Digital Formant Vocoder, *IEEE Transactions on Communications*, Volume COM-26, No.3, March 1978, pp. 344–355.
80. F. Itakura and S. Saito, Analysis Synthesis Telephony based on the Maximum Likelihood Method, *Proceedings of the Sixth Inter-*

national Congress on Acoustics, Tokyo, 1968, Paper C-5-5, pp. C17–20.

81. B. Atal and S.L. Hanauer, Speech Analysis and Synthesis by Linear Prediction of the Speech Wave, *The Journal of the Acoustical Society of America*, Volume 50, August 1971, pp. 737–655.
82. J.D. Markel and A.H. Gray Jr., A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-22, April 1974, pp. 124–134.
83. F. Itakura and S. Saito, On the Optimum Quantization of Feature Parameters in the PARCOR Speech Synthesizer, *IEEE Conference on Speech Communications and Processing*, New York, April 1972, pp. 434–437.
84. J.R. Haskew, J.M. Kelly, R.M. Kelly and T.H. McKinney, Results of a Study of the Linear Prediction Vocoder, *IEEE Transactions on Communications*, Volume COM-21, No.9, September 1973, pp. 1008–1015.
85. R. Viswanathan and J. Makhoul, Quantization Properties of Transmission Parameters in Linear Predictive Systems, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-23, June 1975, pp. 309–321.
86. A.H. Gray Jr., R.M. Gray and J.D. Markel, Comparison of Optimal Quantizations of Speech Reflection Coefficients, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-25, No.1, February 1977, pp. 9–23.
87. J.D. Markel and A.H. Gray Jr., Implementation and Comparison of Two Transformed Reflection Coefficient Scalar Quantization Methods, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-28, No.5, October 1980, pp. 575–583.
88. P. Kroon, *Time-Domain Coding of (near) Toll Quality Speech at Rates Below 16 Kb/s*, Thesis, Delft University of Technology, Delft, The Netherlands, 1985, Section 6.4: Quantizing And Encoding The Filter Parameters.
89. F. Itakura, Line Spectrum Representation of Linear Predictor Coefficients of Speech Signals, *The Journal of the Acoustical Society of America*, Volume 57, Supplement No.1, 1975, p.S35 (Abstract).
90. K.K. Soong and B-H. Juang, Line Spectrum Pair (LSP) And Speech Data Compression, *IEEE International Conference on*

Acoustics, Speech and Signal Processing, San Diego, 1984, pp. 1.10.1–1.10.4.

91. P. Kabal and R.P. Ramachandran, The Computation of Line Spectral Frequencies Using Chebychev Polynomials, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-34, No.6, December 1986, pp. 1419–1426.
92. G.S. Kang and L.J. Fransen, Application of Line-Spectrum Pairs to Low-Bit-Rate Speech Encoders, *IEEE International Conference on Acoustics, Speech and Signal Processing*, New York, 1985, pp. 244–247.
93. M.R. Sambur, A.E. Rosenberg, L.R.Rabiner and C.A. McGonegal, On reducing the buzz in LPC synthesis, *The Journal of the Acoustical Society of America*, Volume 63, No.3, March 1978, pp. 918–924.
94. J.M. Makhoul, R. Viswanathan, R. Schwartz and A.W.F. Huggins, A mixed-source model for speech compression and synthesis, *The Journal of the Acoustical Society of America*, Volume 64, No.6, December 1978, pp. 1577–1581.
95. P. Hedelin, High Quality Glottal LPC-Vocoding, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, 1986, pp. 465–468.
96. S. Maitra and C.R. Davis, A Speech Digitizer at 2400 Bits/s, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No.6, December 1979, pp. 729–733.
97. A. van Leeuwaarden and R.J. Sluijter, An LPC Vocoder for Transmission of Speech at 2400 Bit/s, *Philips Research Laboratories Internal Report*, No.5724, March 1982.
98. T.E. Tremain, The Government Standard Linear Predictive Coding Algorithm: LPC-10, *Speech Technology*, Volume 1, No.2, April 1982, pp. 40–48.
99. J.A. Feldman, E.M. Hofstetter and M.L. Malpass, A Compact, Flexible LPC Vocoder Based on a Commercial Signal Processing Microcomputer, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-31, No.1, February 1983, pp. 252–257.
100. *μPD7720 Signal Processing Interface (SPI)*, Product Description, NEC Electronics (Europe) GmbH, Düsseldorf, Germany, July 1981.

101. M. Grützmacher and W. Lottermoser, Über ein Verfahren zur trägheitsfreien Aufzeichnung von Melodiekurven, *Akust. Z.*, Band 2, September 1937, pp. 242–248.
102. H. Dudley, Remaking Speech, *The Journal of the Acoustical Society of America*, Volume 11, No.2, October 1939, pp. 169–177.
103. O.O. Gruenz Jr. and L.O. Schot, Extraction and Portrayal of Pitch of Speech Sounds, *The Journal of the Acoustical Society of America*, Volume 21, No.5, September 1949, pp. 487–495.
104. B. Gold, Computer Program for Pitch Extraction, *The Journal of the Acoustical Society of America*, Volume 34, No.7, July 1962, pp. 916–921.
105. B. Gold, Note on Buzz-Hiss Detection, *The Journal of the Acoustical Society of America*, Volume 36, No.9, September 1964, pp. 1659–1661.
106. B. Gold and L. Rabiner, Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain, *The Journal of the Acoustical Society of America*, Volume 46, No.2 (Part 2), 1969, pp. 442–448.
107. R.L. Miller and E.S. Weibel, Measurement of the Fundamental Period of Speech Using a Delay Line, *The Journal of the Acoustical Society of America*, Volume 28, No.4, July 1956, p.761 (Abstract).
108. J.S. Gill, Automatic Extraction of the Excitation Function of Speech with Particular Reference to the Use of Correlation Methods, 3^{rd} *International Congress on Acoustics ICA-59*, Stuttgart, 1959, pp. 217–220.
109. M.M. Sondhi, New Methods of Pitch Extraction, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-16, June 1968, pp. 262–266.
110. A.M. Noll, Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection, *The Journal of the Acoustical Society of America*, Volume 36, No.2, February 1964, pp. 296–302.
111. J.D. Markel, The SIFT Algorithm for Fundamental Frequency Estimation, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-20, No.5, December 1972, pp. 367–377.
112. P. Butler and D.J.H. Moore, Pitch Detection in Speech, *Australian Telecommunication Research (A.T.R.)*, Volume 7, No.2, 1973, pp. 39–46.
113. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg and H.J. Manley, Average Magnitude Difference Function Pitch Extractor, *IEEE*

Transactions on Acoustics, Speech, and Signal Processing, Volume ASSP-22, No.5, October 1974, pp. 353–362.

114. C.K. Un and S-C. Yang, A Pitch Extraction Algorithm Based on LPC Inverse Filtering and AMDF, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-25, No.6, December 1977, pp. 565–572.
115. L.R. Rabiner, M.J. Cheng, A.E. Rosenberg and C.A. McGonegal, A Comparative Performance Study of Several Pitch Detection Algorithms, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-24, October 1976, pp. 399–417.
116. C.A. McGonegal, L.R. Rabiner, and A.E. Rosenberg, A Subjective Evaluation of Pitch Detection Methods Using LPC Synthesized Speech, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-25, No.3, June 1977, pp. 221–229.
117. R.J. Sluijter, H.J. Kotmans and A. van Leeuwaarden, The Design of a Real-Time "Harmonic Sieve" Pitch Extractor, *Philips Research Laboratories Internal Report*, No.5526, October 1979.
118. R.J. Sluijter, H.J. Kotmans and A. van Leeuwaarden, The Harmonic-Sieve Method for Pitch Extraction From Speech and a Hardware Model Applicable to Vocoder Systems, *Proceedings of the International Zurich Seminar on Digital Communications*, March 1980, pp. E2.1–E2.6.
119. R.J. Sluijter, H.J. Kotmans and A. van Leeuwaarden, A Novel Method for Pitch Extraction from Speech and a Hardware Model Applicable to Vocoder Systems, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Denver, 1980, pp. 45–48.
120. R.J. Sluijter, Redesign of the Harmonic-Sieve Pitch Extractor and an Appropriate Voiced-Unvoiced Detector, *Philips Research Laboratories Internal Report*, TN 198/81, October 1981.
121. R.J. Sluijter, H.J. Kotmans and T.A.C.M. Claasen, Improvements of the Harmonic-Sieve Pitch Extraction Scheme and an Appropriate Method for Voiced-Unvoiced Detection, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 188–191.
122. H. Duifhuis, L.F. Willems and R.J. Sluijter, Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception, *The Journal of the Acoustical Society of America*, Volume 71, No.6, June 1982, pp. 1568–1580.

123. W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.
124. L.R. Rabiner, M.R. Sambur and C.E. Schmidt, Applications of a Nonlinear Smoothing Algorithm to Speech Processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-23, No.6, December 1975, pp. 552–557.
125. B.G. Secrest and G.R. Doddington, Post-processing Techniques for Voice Pitch Trackers, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 172–175.
126. B.G. Secrest and G.R. Doddington, An Integrated Pitch Tracking Algorithm for Speech Systems, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 1352–1355.
127. C.B.H. Feldman, Band Compression System, *USA Patent*, No. 2 817 711, 1957, filed 10 May, 1954
128. M.R. Schroeder and E.E. David, A Vocoder for Transmitting 10 kc/s Speech over a 3.5 kc/s Channel, *Acustica*, Volume 10, 1960, pp. 35–43.
129. J.L. Flanagan, A Resonance-Vocoder and Baseband Complement: A Hybrid System for Speech Transmission, *IRE Transactions on Audio*, May-June 1960, pp. 95–102.
130. E.E. David Jr., M.R. Schroeder, B.F. Logan and A.J. Prestigiacomo, Voice-Excited Vocoders for Practical Speech Bandwidth Reduction, *IRE Transactions on Information Theory*, Volume IT-8, September 1962, pp. S101–S105.
131. R.M. Golden, Digital Computer Simulation of a Sampled-Data Voice-Excited Vocoder, *The Journal of the Acoustical Society of America*, Volume 35, No.9, September 1963, pp. 1358–1366.
132. B. Gold and J. Tierney, Digitized Voice-Excited Vocoder for Telephone-Quality Inputs, Using Bandpass Sampling of the Baseband Signal, *The Journal of the Acoustical Society of America*, Volume 37, April 1965, pp. 753–754.
133. J-F. Zurcher, P. Graillot and M. Cartier, Speech Digitalization with Channel Vocoders, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, 1976, pp. 95–98.
134. R.J. Sluijter, The Volksvocoder and its Improved Version, *Philips Research Laboratories Internal Report*, No. 5142, December 1975.
135. R.J. Sluijter, Eenvoudige Vocoder voor Digitale Spraakoverdracht, *Tijdschrift van het Nederlands Elektronica- en Radiogenootschap*,

Deel 41, Nr.3, 1976, pp. 67–72. Also in: *Voordrachten over Diverse Akoestische Onderwerpen*, Nederlands Akoestisch Genootschap, Publikatie Nr. 40, april 1977, pp. 23–32.

136. R.J. Sluijter, The Volksvocoder: Circuit Diagrams, *Philips Research Laboratories Internal Report*, TN 5/77, Januari 1977.
137. R.J. Sluijter, Design Tables for RC Active Cauer Filters, *Journal of Applied Science and Engineering A*, 3, 1978, pp. 1–14.
138. J.G. Dunn, An Experimental 9600-bits/s Voice Digitizer Employing Adaptive Prediction, *IEEE Transactions on Communication Technology*, Volume COM-19, No.6, December 1971, pp. 1021–1032.
139. C.K. Un and D.T. Magill, The Residual-Excited Linear Prediction Vocoder with Transmission Rate Below 9.6 kbit/s, *IEEE Transactions on Communication*, Volume COM-23, No.12, December 1975, pp. 1466–1473.
140. B.S. Atal, M.R. Schroeder and V. Stover, Voice-Excited Predictive Coding System for Low Bit-Rate Transmission of Speech, *IEEE International Conference on Communications (ICC)*, San Francisco, 1975, pp. (30-37)–(30-40).
141. D. Esteban, C. Galand, D. Mauduit and J. Menez, 9.6/7.2 KBPS Voice Excited Predictive Coder (VEPC), *IEEE International Conference on Acoustics, Speech and Signal Processing*, Tulsa, 1978, pp. 307–311.
142. J. Makhoul and M. Berouti, High-Frequency Regeneration in Speech Coding Systems, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Washington, 1979, pp. 428–431.
143. H. Katterfeldt, A DFT-Based Residual-Excited Linear Predictive Coder (RELP) for 4.8 and 9.6 kb/s, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, 1981, pp. 824–827.
144. H. Katterfeldt and E. Behl, Implementation of a Robust RELP Speech Coder, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 1316–1319.
145. A.B. Carlson, *Communication Systems*, McGraw-Hill Book Company, New York, 1986
146. C.K. Un and J.R. Lee, On Spectral Flattening Techniques in Residual-Excited Linear Prediction Vocoding, *IEEE International*

Conference on Acoustics, Speech and Signal Processing, Paris, 1982, pp. 216–219.

147. V.R. Viswanathan, A.L. Higgins and W.H. Russell, Design of a Robust Baseband LPC Coder for Speech Transmission Over 9.6 Kbit/s Noisy Channels, *IEEE Transactions on Communication*, Volume COM-30, No.4, April 1982, pp. 663–673.
148. M.M. Arjmand and G.R. Doddington, Pitch-Congruent Baseband Speech Coding, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 1324–1327.
149. P. Hedelin, RELP-Vocoding with Uniform and Non-Uniform Down-Sampling, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Boston, 1983, pp. 1320–1323.
150. R.J. Sluijter, G.J. Bosscha and H.M.P.T. Schmitz, A 9.6 Kbit/s Speech Coder for Mobile Radio Applications, *IEEE International Conference on Communications (ICC)*, Amsterdam, 1984, pp. 1159–1162. (P. Dewilde and C.A. May (Eds.), *Links for the Future*, IEEE/ Elsevier Science Publishers B.V., North Holland, 1984.)
151. R.J. Sluijter, Digital Speech Coder with Baseband Residual Coding, *USA patent 4,752,956*, June 1988 (first filed in March 1984).
152. P. Vary and R.J. Sluijter, Speech Processing in the Mobile Radio Terminal, *Nordic Seminar on Digital Land Mobile Radio Communication*, Espoo, Finland, February 1985, pp. 67–76.
153. R.J. Sluijter, Spraakcodering voor het Mobiele Radiokanaal, *K.I.V.I. Leergang*, T.H.E. (Technical University Eindhoven), Mobiele Communicatie, Eindhoven, April 1983.
154. R.J. Sluijter, Digitization of Speech, *Philips Technical Review*, Volume 41, No.7/8, 1983/1984, pp. 201–223. (Dutch version: Digitalisering van Spraak, *Philips Technisch Tijdschrift*, Jaargang 41, No.7/8, 1983.)
155. R.J. Sluijter, The state of the art in speech coding, *Proceedings of the 9th European Solid-State Circuits Conference*, Lausanne, Switserland, September 1983 (ESSCIRC'83), pp. 33–40.
156. B.S. Lee, H.H. Lee, B.C. Shin and H.S. Lee, Implementation of a Multirate Speech Digitizer, *IEEE Transactions on Communication*, Volume COM-31, No.6, June 1983, pp. 775–783.
157. M. Dankberg, R. Iltis, D. Saxton and P. Wilson, Implementation of the RELP Vocoder Using the TMS320, *IEEE International Con-*

ference on Acoustics, Speech and Signal Processing, San Diego, 1984, pp. 27.8.1–27.8.4.

158. C.G. Bell, H. Fujisaki, J.M. Heinz, K.N. Stevens and A.S. House, Reduction of Speech Spectra by Analysis-by-Synthesis Techniques, *The Journal of the Acoustical Society of America*, Volume 33, No.12, December 1961, pp. 1725–1736.
159. M.R. Schroeder and B.S. Atal, Rate Distortion Theory and Predictive Coding, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Atlanta, 1981, pp. 201–204.
160. B.S. Atal and J.R. Remde, Digital Speech Coder, *USA Patent* 4 472 832, 1984 (filed December 1981).
161. B.S. Atal, Predictive Coding of Speech at Low Bit Rates, *IEEE Transactions on Communication*, Volume COM-30, No.4, April 1982, pp. 600–614.
162. P. Kroon, E.F.A. Deprettere, R.J. Sluijter, Multi-Pulse Excitation Linear-Predictive Speech Coder, *USA Patent* 4 932 061, 1990 (first filed March 1982)
163. E.F. Deprettere and P. Kroon, Regular Excitation Reduction for Effective and Efficient LP-Coding of Speech, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Tampa, 1985, pp. 965–968.
164. B.S. Atal and J.R. Remde, A new Model of LPC Excitation for Producing Natural-Sounding Speech at Low Bit Rates, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Paris, 1982, pp. 614–617.
165. S. Singhal and B.S. Atal, Improving Performance of Multi-pulse LPC Coders at Low Bit Rates, *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, 1984, pp. 1.3.1–1.3.4.
166. P. Kroon and E.F. Deprettere, Experimental Evaluation of Different Approaches to the Multi-Pulse Coder, *IEEE International Conference on Acoustics, Speech and Signal Processing*, San Diego, 1984, pp. 10.4.1–10.4.4.
167. B.S. Atal and M.R. Schroeder, Stochastic Coding of Speech Signals at Very Low Bit Rates, *IEEE International Conference on Communications (ICC)*, Amsterdam, 1984, pp. 1610–1613. (P. Dewilde and C.A. May (Eds.), *Links for the Future*, IEEE/Elsevier Science Publishers B.V., North Holland, 1984.)

168. M.R. Schroeder and B.S. Atal, Code-Excited Linear Prediction (CELP): High-Quality Speech at Very Low Bit Rates, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Tampa, 1985, pp. 937–940.

169. P. Kroon, *Time-Domain Coding of (Near) Toll Quality Speech at Rates Below 16 Kb/s*, Ph.D. Thesis, Technical University of Delft, The Netherlands, May 1985.

170. J.P. Adoul, F. Didelot, P. Mabilleau and S. Morisette, Generalization of the Multi-pulse Coding for Low Bit Rate Coding Purposes: The Generalized Decimation, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Tampa, 1985, pp. 256–259.

171. R.W. Schafer and L.R. Rabiner, Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis, *IEEE Transactions on Audio and Electroacoustics*, Volume AU-21, No.3, June 1973, pp. 165–174.

172. R.E. Crochiere, S.A. Webber and J.L. Flanagan, Digital Coding of Speech in Sub-Bands, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Philadelphia, 1976, pp. 233–236.

173. J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E Crochiere, N.S. Jayant and J.M. Tribolet, Speech Coding, *IEEE Transactions on Communication*, Volume COM-27, No.4, April 1979, pp. 710–736.

174. J.M. Tribolet and R.E. Crochiere, Frequency Domain Coding of Speech, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-27, No.5, October 1979, pp. 512–530.

175. R.E. Crochiere and L.R. Rabiner, *Multi-rate Digital Signal Processing*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey, 1983.

176. J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, Berlin, 1976.

177. J.L. Flanagan and R.M. Golden, Phase Vocoder, *The Bell System Technical Journal*, November 1966, pp. 1493–1509.

178. M.R. Portnoff, Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Volume ASSP-24, No.3, June 1976, pp. 243–248.

Chapter 4

Design of the GSM speech codec

In this chapter, which constitutes the core of this thesis, the design and architectural features of the GSM full-rate standard speech coder are discussed. The design of this coder stems from the mid eighties and the standard was completed in 1988. The techniques used in the design are built on the fundamentals described in the previous chapters, including quantization techniques, the composite acoustic tube model, LPC (linear prediction coding), LTP (long-term prediction) and pitch detection, RELP (residual-excited linear-prediction) coding and the MPE (multi-pulse excitation), RPE (regular-pulse excitation) and CELP (code-excited linear prediction) methods of analysis-by-synthesis coding. At the time of design, it had been recognized that the analysis-by-synthesis coding paradigms could provide the best possible performance with respect to speech quality and robustness, but that the complexity was too high to accommodate low-power real-time implementation as required by the mobile application. Straightforward application of the RELP coding paradigm could meet the complexity requirements, but it had some other design problems. These concern spectral folding of high-pass filtered speech ("telephone speech") as explained in the final paragraph of Section 3.4.2 and the design of the low-pass filter as discussed in Section 3.4.3.

The thread of this chapter concerns the quest for low-complexity analysis-by-synthesis solutions, which at the same time maintain their original performance as much as possible. At the start of the design activities, the complexity of the CELP method was far beyond the re-

quirements and it was not yet understood how to deal with this problem, so that the focus was on MPE and RPE methods. Pulse-amplitude computation in MPE and RPE coders included solving sets of equations (typically of order 10–16) two hundred times a second, which accounted for the major part of the complexity. In the next sections it is explained how we solved the complexity issue by innovative explicit-form designs, thus avoiding solving these sets of equations.

The chapter is organized as follows. In the first section of this chapter, the analysis-by-synthesis system impulse response and the basic algorithms of MPE and RPE using this system impulse response, are specified in detail. Section 4.2 recapitulates some strategies to achieve low complexity which were already known at the beginning of the design period, including the autocorrelation method, presetting the system impulse response and truncation of that impulse response. None of these approaches was sophisticated enough to provide the required solution, however. In Section 4.3 the precise time- and frequency-domain, as well as the perceptual, properties of the system impulse response are analysed and characterised. As a result, the concept of "radical truncation" of the system impulse response is introduced and the impact on MPE and RPE coding methods is investigated. The combination with the autocorrelation method fully exploits the opportunity offered by radical truncation and it returns an attractive explicit-form solution. Alternatively, a preset fixed system impulse response adjusted to the long-term spectral properties of speech is investigated. It is found that combining presetting with RPE and tuning the preset impulse response to the particular decimation factor of the RPE design, gives rise to an even more attractive explicit-form solution. This solution yields a low-complexity RELP-like architecture, but without the RELP problems. The low-pass filter has been replaced by a "block-wise convolver" with an implicit design which maintains the optimum analysis-by-synthesis pulse amplitudes, given the constraints on the system impulse response. The rigid down-sampling and spectral folding grid has been replaced by the more sophisticated dynamic RPE grid selection. Section 4.4 deals with the inclusion of LTP into this architecture to form the well-known LPC-LTP-RPE cascade. Finally, the selection process of this coder as the full-rate GSM standard speech coder, at a bit rate of 13 kbit/s, and the performance of this standard coder in the mobile environment, are recapitulated in Section 4.5.

For the first time in open literature, the underlying design alter-

natives from which this coder has evolved are discussed in detail here. Several of these details have only been described in the proprietary report [11] and in the original European and US patents [12], and some details have never been published at all. For instance, a distinction is made (in Section 4.3.8) between explicit "by approximation" and explicit "by construction" for the first time and it is explained (in Sections 4.3.5 and 4.3.7) how the coefficients of the convolver are determined to obtain an explicit-by-construction RPE system for a decimation factor of three. This decimation factor and the produced convolver coefficients are actually used in the standard coder. Another unpublished issue is the way how the common-practice inclusion of LTP in state-of-the-art linear prediction based analysis-by-synthesis architectures according to Figure 3.20, relates to the GSM-standard-coder architecture. This is explained in Section 4.4.1.

4.1 The MPE and RPE methods in detail

The signal path of the analysis-by-synthesis loop of Figure 3.19 can be represented by

$$E(z) = W(z)\left\{S(z) - \frac{X(z)}{A(z)}\right\} , \tag{4.1}$$

where $S(z)$, $X(z)$ and $E(z)$ stand for the z-transforms of the input speech signal $s[n]$, the excitation signal $x[n]$ and the perceptually weighted error signal $e[n]$, respectively. The transfer function $W(z)$ stands for the perceptual weighting filter and it is defined by Equation 3.43. As usual, $A(z)$ represents the LPC inverse filter. By rearranging the terms according to

$$E(z) = \{S(z)A(z) - X(z)\}\frac{1}{A(z/\gamma)} , \tag{4.2}$$

the equivalent circuit of Figure 4.1 is obtained. The forward adaptive filter $A(z)$ generates a continuous stream of prediction residual samples $d[n]$ from the speech signal $s[n]$. The prediction coefficients are determined from segments of $s[n]$, usually every 20 ms, and the length of a segment may be longer so that successive segments overlap. As will become apparent, the explicit presence of the prediction residual in this system configuration contributes to the computational efficiency of the eventual system. The system transfer function $H(z)$ from the excitation source $x[n]$ to the error signal $e[n]$ is again $H(z) = 1/A(z/\gamma)$ with impulse response $h[n]$.

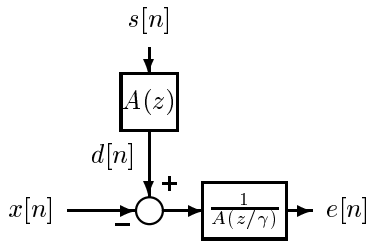

Figure 4.1: Basic system configuration representing Figure 3.19 according to Equation 4.2.

Although the GSM full-rate coder is based on RPE and LTP ("long-term prediction" or pitch prediction) we will include the MPE option until it becomes clear that RPE is the preferred choice. Only MPE with sequential search and RPE will be considered because an exhaustive search for either MPE or CELP was far beyond the affordable complexity of the intended application in mobile telephony. Effective alternatives for the exhaustive search in CELP were not yet known in the early 1980s. No attention will be paid to LTP, as yet. It will be discussed and included into the final architecture in Section 4.4.

In the basic MPE and RPE methods consecutive (rectangular) windows with the length of a subsegment of typically 5 ms are applied to the error sequence $e[n]$. Each window ranges from 0 until $N-1$ and equals one within the subsegment and zero outside. The sampling frequency is always 8 kHz, unless specified otherwise.

4.1.1 The sequential MPE algorithm

The subsequent steps for each subsegment to obtain K optimised pulse positions n_k and amplitudes $x[n_k]$ according to the sequential MPE algorithm as explained in Section 3.5.1 are the following.

1. Determine the initial error sequence $e_0[n]$ for the current subsegment according to Equation 3.45, by feeding the speech signal $s[n]$ through the system with zero excitation $x[n] = 0$, $n = 0, 1, \ldots, N-1$. The initial states of the filters $A(z)$ and $H(z)$ depend on the history of excitation signals in previous subsegments.

2. Determine the impulse response $h[n]$ and repeat the next steps a, b and c for $k = 1, 2, \ldots, K$:

a. determine the location n_k of the k^{th} pulse on the basis of Equation 3.50 by the search over $i = 0, 1, \ldots, N-1$ for

$$n_k = \arg\max_i \frac{\left(\sum_{n=0}^{N-1} e_{k-1}[n]h[n-i]\right)^2}{\sum_{n=0}^{N-1} h^2[n-i]} \; ; \tag{4.3}$$

b. compute the associated pulse amplitude on the basis of Equation 3.49 by

$$x[n_k] = \frac{\sum_{n=0}^{N-1} e_{k-1}[n]h[n-n_k]}{\sum_{n=0}^{N-1} h^2[n-n_k]} \; ; \tag{4.4}$$

c. determine the reduced error sequence $e_k[n]$ on the basis of Equation 3.46 by

$$e_k[n] = e_{k-1}[n] - x[n_k]h[n-n_k] \; . \tag{4.5}$$

3. Reinstall the initial state of $H(z)$ and apply the difference between $d[n]$ and the newly found excitation sequence $x[n]$,

$$x[n] = \begin{cases} x[n_k], & n = n_k, \; k = 1, 2. \ldots, K \\ 0, & \text{elsewhere} \end{cases} \; , \tag{4.6}$$

to the filter $H(z)$ to determine its state at the end of the subsegment. This is the initial state of the next subsegment.

As explained in Section 3.5.1 the pulse amplitudes can be recomputed once all pulse locations are known, giving a joint optimisation of all pulse amplitudes. For this purpose, step 2 can be extended in the following way. Representing the K amplitudes by the column vector $\mathbf{x}$ with elements $x[n_k]$, they are implicitly given by

$$\mathbf{C}\mathbf{x} = \mathbf{v} \; , \tag{4.7}$$

where $\mathbf{C}$ is the symmetric $K \times K$ matrix with elements

$$c[n_i, n_k] = \sum_{n=0}^{N-1} h[n-n_i]h[n-n_k] \; , \tag{4.8}$$

and the K elements of the column vector $\mathbf{v}$ are given by

$$v[n_k] = \sum_{n=0}^{N-1} e_0[n]h[n - n_k] \; . \tag{4.9}$$

The recomputed amplitudes are efficiently obtained by solving Equation 4.7 with the aid of the modified Cholesky algorithm.

In the refined re-computation method, also explained in Section 3.5.1, the joint optimisation is applied after each newly found pulse location. For this purpose, the above re-computation method is incorporated within the iteration loop of step 2.

4.1.2 The RPE algorithm

In RPE as described in Section 3.5.3, the $K = N/D$ pulse locations (assuming that N/D is integer) in each subsegment are always given by Equation 3.60 for each grid position $0 \leq \phi \leq D-1$, where D is the downsampling factor which is also often referred to as the decimation factor. The optimised grid position and pulse amplitudes for each subsegment are obtained by the following steps [1].

1. Determine the initial error sequence $e_0[n]$ in the same way as above for MPE.

2. Determine the impulse response $h[n]$ and compute the coefficients of the D symmetric $K \times K$ covariance matrices $\mathbf{C}_\phi$ according to Equation 3.53 by

$$c_\phi[n_i, n_k] = \sum_{n=0}^{N-1} h[n - \phi - (i-1)D]h[n - \phi - (k-1)D] \; , \tag{4.10}$$

   where both i and k range from 1 to K.

3. Compute the elements of the D column vectors $\mathbf{v}_\phi$,

$$v_\phi[n_k] = \sum_{n=0}^{N-1} e_0[n]h[n - \phi - (k-1)D] \; , \tag{4.11}$$

   each for $k = 1, 2, \ldots, K$.

4. Solve the excitation vectors $\mathbf{x}_\phi$ for the D grid positions from

$$\mathbf{C}_\phi \mathbf{x}_\phi = \mathbf{v}_\phi \; , \tag{4.12}$$

   using the modified Cholesky algorithm.

5. Select the optimum grid position p giving the minimum error on the basis of Equation 3.55:

$$p = \arg\max_{\phi} \sum_{k=1}^{K} x_\phi[n_k] v_\phi[n_k] \ . \tag{4.13}$$

6. The excitation sequence $x[n]$ at the selected locations is now given by

$$\left.\begin{array}{rcl} n_k & = & p + (k-1)D \\ x[n_k] & = & x_p[n_k] \end{array}\right\} k = 1, 2, \ldots, K \ , \tag{4.14}$$

and $x[n] = 0$ elsewhere. This sequence in used in the determination of the final state of $H(z)$ for the initiation of the next subsegment, just as described above for the MPE case.

The next sections will be devoted to the exploration of computationally efficient variants of these algorithms in order to enable implementation of an MPE or RPE codec on a single-chip digital signal processor of the early 1980s. The state of the art of such devices in those days is well reflected in [2, 3, 4].

4.2 Early efficient methods

In the construction of the matrices, recursions can be utilized. The matrix equations are very similar to those of the LPC covariance method as described in Section 2.5.1. Accordingly, an "autocorrelation" method was defined in 1984, with the aim to further enhance the computational efficiency in MPE. We will also evaluate this approach when applied to RPE. Other early considered efficiency enhancing measures which will be discussed were truncation of the system impulse response $h[n]$ and making $h[n]$ fixed.

4.2.1 Recursive matrix construction

The coefficients $c[n_i, n_k]$ of the matrices $\mathbf{C}_\phi$ can efficiently be obtained in a recursive way by evaluating them along (sub)diagonals [6] according to

$$c[i-1, k-1] = c[i, k] + h[N-i]h[N-k] \ . \tag{4.15}$$

The computational savings in RPE are evident from Equation 4.10 since interleaving of the matrices $\mathbf{C}_\phi$ yields (sub)diagonals with consecutive coefficients in the recursion. The savings in MPE re-computation

depend on the actual pulse positions, however, and the required matrix coefficients are generally not consecutive in the recursion.

4.2.2 The autocorrelation method

The autocorrelation method was first proposed by Berouti, Garten, Kabal and Mermelstein in 1984 [6]. In the same way as in LPC, the autocorrelation method is obtained by applying a finite-duration window to the data, instead of the error signal, and by including the entire non-zero error sequence. In this particular case, the prediction residual $d[n]$ and the excitation signal $x[n]$, with reference to Figure 4.1, are both windowed by the rectangular N-point window. One consequence of this approach is that, because there was no data for $n < 0$, the filter $H(z) = 1/A(z/\gamma)$ is, for each subsegment, initially supposed to be in zero-state without any hangover from the past. The total square error after applying K pulses is now given by

$$E_K = \sum_{n=0}^{\infty} \left(e_0[n] - \sum_{k=1}^{K} x[n_k] h[n - n_k] \right)^2 , \tag{4.16}$$

where $e_0[n]$ is now defined by

$$e_0[n] = e[n]\big|_{x[n]=0}, \quad n = 0, 1, \ldots, \infty. \tag{4.17}$$

This error sequence $e_0[n]$ consists of the response of the filter $H(z)$ to the windowed portion of the signal $d[n]$, and $e_0[n]$ equals zero for $n < 0$. The response lasts very long, in principle, because $H(z)$ has the infinite impulse response $h[n]$. Putting the partial derivatives of E_K with respect to $x[n_k]$ for all $k = 1, 2, \ldots, K$ to zero, yields

$$\sum_{i=1}^{K} x[n_i] \rho[|n_i - n_k|] = \nu[n_k], \quad k = 1, 2, \ldots, K , \tag{4.18}$$

where

$$\rho[|n_i - n_k|] = \sum_{n=0}^{\infty} h[n] h[n + |n_i - n_k|] , \tag{4.19}$$

and

$$\nu[n_k] = \sum_{n=0}^{\infty} e_0[n] h[n - n_k] . \tag{4.20}$$

The residual error is, in accordance with Equation 3.55, given by

$$E_K = E_0 - \sum_{k=1}^{K} x[n_k]\nu[n_k] \; . \tag{4.21}$$

If we write the above Equation 4.18 as $\mathbf{Rx} = \boldsymbol{\nu}$, then $\mathbf{R}$ is a symmetrical $K \times K$ matrix with the same entries $\rho[0]$ on the main diagonal. In addition, depending on the particular pulse positions, other entries in the matrix may occur multiple times as well, so that the autocorrelation method requires less computations in the construction of the matrix. Moreover, separate treatment of the history of the filters is avoided.

In the sequential MPE search a single pulse is searched in accordance with Equations 4.3 and 4.4. For the k^{th} pulse using the autocorrelation method the first equation reduces to

$$n_k = \arg\max_i \frac{1}{\rho[0]} \left(\sum_{n=0}^{\infty} e_{k-1}[n]h[n-i] \right)^2 \; , \tag{4.22}$$

which can further be simplified by omitting $\rho[0]$ and replacing the square by the absolute value. The second equation reduces to

$$x[n_k] = \frac{1}{\rho[0]} \sum_{n=0}^{\infty} e_{k-1}[n]h[n-n_k] \; . \tag{4.23}$$

The total savings resulting from the autocorrelation method for MPE are modest, however, since the complete sequential search, including updating of the error sequence for each newly found pulse, still has to be carried out. Berouti et al. [6] evaluated the speech quality using the autocorrelation method as "...the covariance approach tends to give very slightly better overall SNR values. However, listening tests indicate that differences are small and not consistently in favour of one method or the other."

For RPE, the computational savings by using the autocorrelation method become significant, however. Because the distance between any RPE pulses in a subsegment amounts to a multiple of the decimation factor D, there are only $K = N/D$ different autocorrelation coefficients $\rho[|n_i - n_k|]$, $|n_i - n_k| = 0, D, 2D \ldots, (K-1)D$. In addition, the matrix $\mathbf{R}$ does not depend on the grid position ϕ anymore, since the autocorrelation coefficients only depend on the distance between the pulses and the distance, in turn, only depends on D. Moreover, the matrix becomes a Toeplitz matrix and the equations

$$\mathbf{Rx}_\phi = \boldsymbol{\nu}_\phi \tag{4.24}$$

can now very efficiently be solved by using the Levinson recursion (see Section 2.5.1), which is simpler than the modified Cholesky solution method. A complete analysis still requires the equation to be solved D times, and the selection of the optimum grid position using Equation 4.13 is also required.

4.2.3 Presetting the system impulse response

A proposal, may be the first one, to use a fixed (non-adaptive) version of $H(z) = 1/A(z/\gamma)$ was done by Willems and Nayyar of the Institute of Perception Research in 1983 [5]. They reported on an MPE system where $1/A(z/\gamma)$ was replaced by a first-order recursive filter (integrator) with a coefficient of 0.7. The effect of using this filter on the speech quality was the introduction of a little extra coding noise. In addition, they used truncation of the impulse response after 23 sampling periods (sampling frequency 10 kHz), already. As a result of these measures, the computations were more efficient, according to the report, but it was not specified in which particular way.

By using such a fixed filter, the matrices $\mathbf{C}$ and $\mathbf{R}$ become fixed and their inverses can be pre-calculated. By doing so, the explicit expression for the covariance method becomes

$$\mathbf{x} = \mathbf{C}^{-1}\mathbf{v} \tag{4.25}$$

and for the autocorrelation method

$$\mathbf{x} = \mathbf{R}^{-1}\boldsymbol{\nu}\,. \tag{4.26}$$

These expressions have less computational complexity than the original modified Cholesky or Levinson solution methods. This approach can profitably be applied to both the MPE and RPE coding methods, but we will arrive at even simpler solutions. These make use of special preset fixed forms of the impulse response $h[n]$ and in addition of truncation of it.

4.2.4 Truncation of the system impulse response

In [6] it is also mentioned that the impulse response $h[n]$ of the system transfer function $H(z) = 1/A(z/\gamma)$ dies out rather quickly because of the bandwidth expansion and that therefore the impulse response may be truncated. This can lower the upper summation limits of the equations

for the autocorrelation method from infinity to some more practical value. Also this paper does not elaborate on this issue, however.

Truncation of the system impulse response has been investigated more closely by Senensieb, Milbourn, Lloyd and Warrington and the results of this work were reported in 1984 [7]. They made use of the fact that if the transfer function of the synthesis filter is represented by $H_1(z) = 1/A(z)$, with impulse response $h_1[n]$, then also $H_1(z/\gamma) = 1/A(z/\gamma)$ so that $H(z) = H_1(z/\gamma)$. This means that the filter $1/A(z/\gamma)$ obtained by applying the window γ^n, $n = 0, 1, 2, \ldots$, to the set of a-parameters of the synthesis filter $1/A(z)$, can alternatively be realized by a filter with impulse response

$$h[n] = h_1[n]\gamma^n \; . \tag{4.27}$$

This implies that the filter with impulse response $h[n]$ can be approximated by a filter the impulse response of which is truncated to a length in the order of the subsegment length N. As mentioned earlier, a usual value of γ is 0.8 and a usual subsegment length is 40 samples on an 8 kHz sampling basis. After half a subsegment length the γ-window has been decayed to about one percent and to about 10^{-3} after 31 sampling periods. The approximation can be improved by lowering γ.

In the sequential MPE search the summation limits in Equations 4.3 and 4.4 (covariance method) or Equations 4.22 and 4.23 (autocorrelation method) can be down-sized according to the duration of the truncated impulse response. If the pulses are spaced wider than the duration of the truncated impulse response, the results of the sequential search are optimum within the scope of this constraint (orthogonal) and no recomputation is needed anymore. The paper [7] mentions that such an implementation was realized on the basis of a mean pulse density of 1 pulse every 2 ms. This approach enabled a real-time implementation of an MPE codec at 7.2 kbit/s using an NEC μPD7720 signal processor together with an Intel 8051 microprocessor on a 160 mm $\times$ 233 mm circuit board.

Truncation of the impulse response to a duration in the order of 2 ms, however, does not avoid that in RPE the D sets of equations have to be solved. In order to avoid that, the decimation factor D should equal or exceed the number of samples in that impulse response (16, for 8 kHz sampling frequency), which would be a very impractical value that causes too much loss in speech quality.

4.3 Increasing the efficiency

If, in some way, the condition could be created that the length of the impulse response $h[n]$ does not exceed the pulse spacing, without a significant loss in speech quality, then the computational complexity could be cut dramatically. In that case, the matrix coefficients in $\mathbf{C}$ for the covariance method as well as the matrix coefficients in $\mathbf{R}$ for the autocorrelation method all become zero, except those on the main diagonal, i.e. $c[n_k, n_k]$ and $\rho[0]$, respectively. Consequently, explicit expressions for the pulse amplitudes $x[n_k]$ would be obtained, given by

$$x[n_k] = v[n_k]/c[n_k, n_k] \tag{4.28}$$

for the covariance method and by

$$x[n_k] = \nu[n_k]/\rho[0] \tag{4.29}$$

for the autocorrelation method. These expressions are even simpler than in the case of a preset fixed $h[n]$ mentioned in the previous section. In the next subsections, we will explore ways to enable these conditions. These have also briefly been described in the RPE patent [12]. Before discussing any further truncation and/or preset of $h[n]$, we will have a closer look on $h[n]$ itself.

4.3.1 Characterisation of the system transfer function

The system transfer function $H(z) = 1/A(z/\gamma)$ is, as we have seen before, a smoothed version of $1/A(z)$. Figure 4.2 shows the magnitudes of successive spectra $1/A(z)$ of a portion of speech from the word "coaches" as used in Section 2.1. This portion concerns 21 spectra around the transition from the last vowel "e" into the final consonant "s". The sampling frequency is 8 kHz. The spectra are computed from 20-ms segments which are Hamming windowed and subsequently subject to a tenth-order LPC analysis using the autocorrelation method. Adjacent spectra are obtained by shifting the time window over 10 ms. Clearly visible in Figure 4.2 is the evolution from the vowel spectra with their steady state formants and with the bulk of the energy in the low frequency region into the unvoiced fricative with prevailing energy in the high frequency region. Figure 4.3 shows bandwidth expanded spectra $|1/A(z/\gamma)|$ with the usual value of $\gamma = 0.8$, for the same speech segments. It can be seen that the spectra have lost their crisp formant

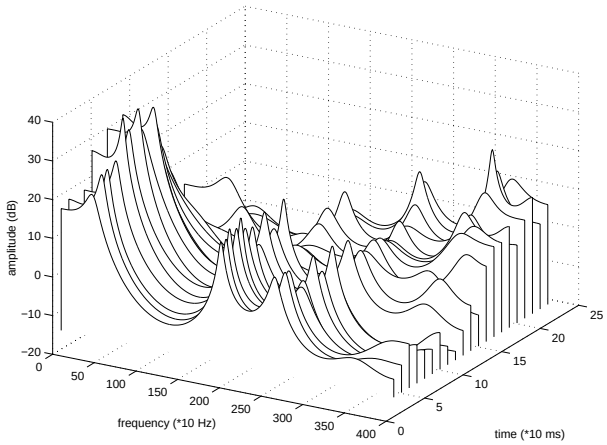

Figure 4.2: $|1/A(e^{j\theta})|$ of a sequence from "coach*es*".

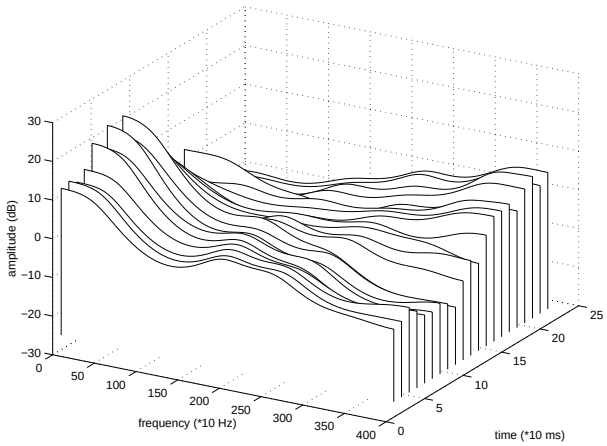

Figure 4.3: $|1/A(e^{j\theta}/0.8)|$ of the same sequence from "coach*es*".

structure and only smooth valleys and hills are left. The general trends or tilts of the various spectra are still there, however, and it is clear that the voiced spectra have retained their decay with frequency and the unvoiced spectra just the opposite. Figures 4.4 and 4.5 show the same but now with $\gamma = 0.7$ and $\gamma = 0.6$, respectively. As compared to the case $\gamma = 0.8$, some extra smoothing has taken place but the general trends stay.

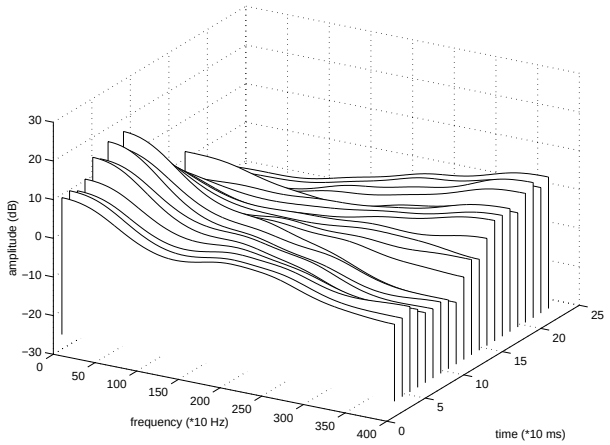

Figure 4.4: $|1/A(e^{j\theta}/0.7)|$ of the same sequence.

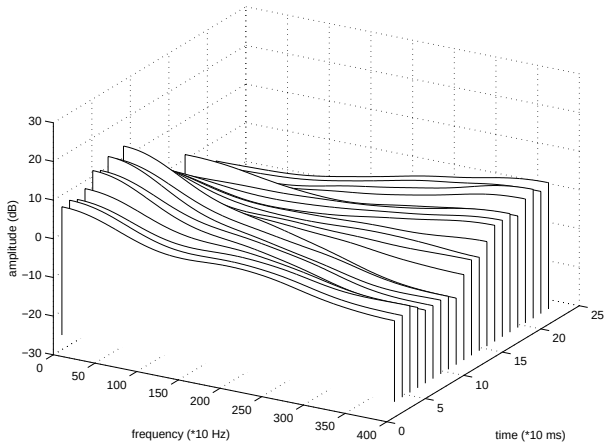

Figure 4.5: $|1/A(e^{j\theta}/0.6)|$ of the same sequence.

The time functions, or impulse responses, belonging to the first and last spectra of the same sequence are shown in Figure 4.6 for $\gamma = 0.8$ and in Figure 4.7 for $\gamma = 0.6$. It is observed that the impulse response is only a few samples long, especially in the case of $\gamma = 0.6$. So, we see that it is possible to truncate the impulse response radically to a duration of a few samples instead of a few tens of samples. The perceptual effect of such a radically truncated impulse response is apparently comparable

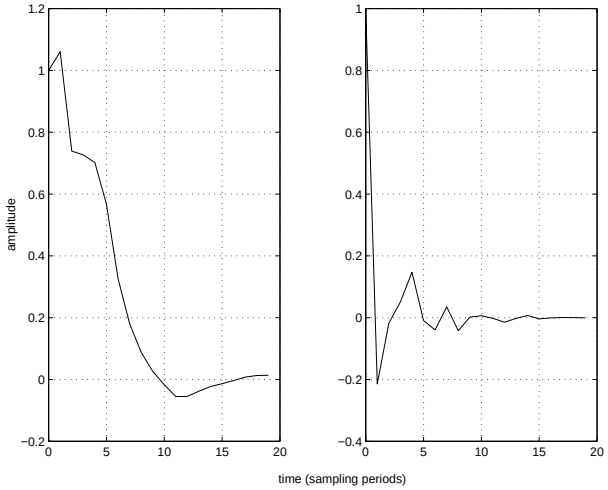

Figure 4.6: Impulse responses of the first (left) and the last (right) transfer functions $1/A(z/0.8)$ of the same sequence.

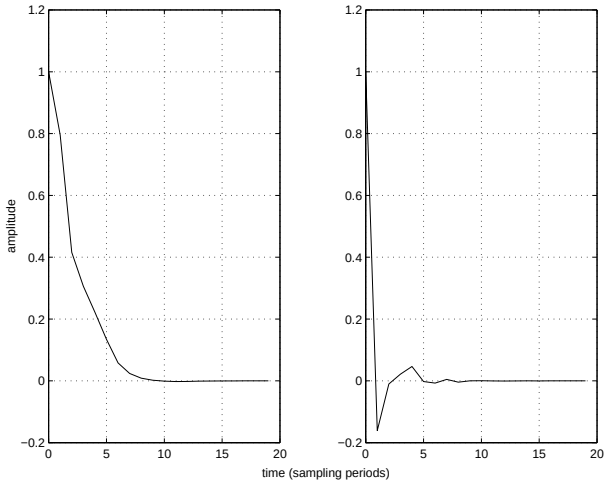

Figure 4.7: Impulse responses of the first (left) and the last (right) transfer functions $1/A(z/0.6)$ of the same sequence.

to the un-truncated impulse response with a γ-value of 0.6 or 0.7. As already stated before in Section 3.5, the perceptual optimum of γ is quite broad and a value of 0.7 or even 0.6 will cause only subtle quality differences as compared to the case $\gamma = 0.8$.

Next, radical truncation of the system impulse response and the consequences of applying this to MPE and RPE are elaborated.

4.3.2 Radical truncation of the impulse response

Good results have been obtained by truncating the impulse response $h_1[n]$ with a triangular window $\lambda[n]$ with a length of $\Lambda = 4$ samples according to

$$\lambda[n] = \begin{cases} 1 - n/\Lambda\ , & n = 0, 1, \ldots, \Lambda - 1 \\ 0\ , & \text{otherwise} \end{cases} , \tag{4.30}$$

instead of the γ-window. In this radical approach the IIR filter with impulse response $h[n]$ is, in agreement with Equation 4.27, replaced by a 4-coefficient FIR filter with impulse response

$$h_\lambda[n] = h_1[n]\lambda[n]\ . \tag{4.31}$$

The magnitudes of the transfer functions $H_\lambda(z)$ of the same sequence of speech segments as above, are shown in Figure 4.8. Comparison of

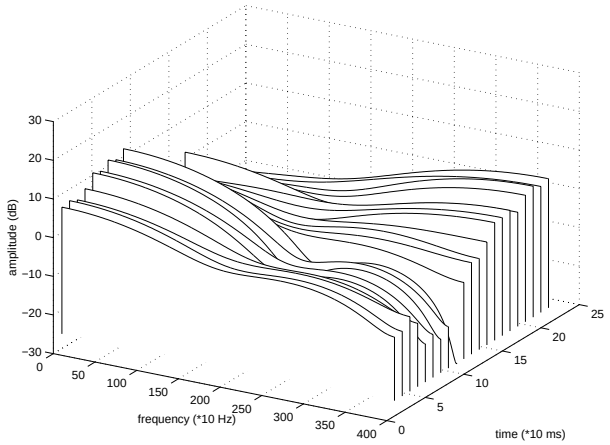

Figure 4.8: Spectra of truncated impulse responses $h_\lambda[n]$ with $\Lambda = 4$ of the same sequence.

these spectra to those of Figure 4.4 and 4.5 reveals that the perceptual differences with the application of $A(z/\gamma)$ with $\gamma = 0.6$ or even 0.7, must be very small. Actual experiments and informal listening tests have confirmed this. The radical truncation brings explicit forms of MPE and RPE if the pulses are spaced by at least Λ sampling periods.

4.3.3 MPE with radical truncation

Let us consider such an MPE coder with radical truncation using the autocorrelation method. In the sequential search, all pulse locations are allowed for the first pulse. For the k^{th} pulse, however, the pulse locations n_i, $i = 1, 2, \ldots, k-1$ have already been allocated so that the pulse locations

$$n = n_i - (\Lambda - 1), \ldots, n_i - 1, n_i, n_i + 1, \ldots, n_i + \Lambda - 1 \tag{4.32}$$

are not allowed in order to maintain the prescribed inter-pulse distances. The k^{th} location is found on the basis of Equation 4.22 by

$$n_k = \arg\max_i \frac{1}{\rho_\lambda[0]} \left| \sum_{n=i}^{i+\Lambda-1} e_{k-1}[n] h_\lambda[n-i] \right| , \tag{4.33}$$

where i is limited to the allowed region and $\rho_\lambda[j]$ is, in agreement with Equation 4.19, defined by

$$\rho_\lambda[j] = \rho_\lambda[|j|] = \sum_{n=0}^{\Lambda-1} h_\lambda[n] h_\lambda[n+|j|], \quad j = -(\Lambda-1), \ldots, \Lambda - 1 \, . \tag{4.34}$$

In the updates of the error sequence according to Equation 4.5 during the first $k-1$ iterations of the sequential search, this allowed region is not touched upon because these updates are limited to the areas given by Equation 4.32. This implies that in the allowed region $e_{k-1}[n] = e_0[n]$ so that n_k can also be found by

$$n_k = \arg\max_i \frac{1}{\rho_\lambda[0]} \left| \sum_{n=i}^{i+\Lambda-1} e_0[n] h_\lambda[n-i] \right| . \tag{4.35}$$

Thus, the amplitude of the k^{th} pulse according to Equation 4.23 can also be written as

$$x[n_k] = \frac{1}{\rho_\lambda[0]} \sum_{n=n_k}^{n_k+\Lambda-1} e_0[n] h_\lambda[n-n_k] \, . \tag{4.36}$$

Since the MPE pulses are mutually independent, re-computation is now redundant.

4.3.4 RPE with radical truncation

In the RPE case with radical truncation and $\Lambda \leq D$, the autocorrelation coefficients according to Equations 4.19 and 4.34 become

$$\rho_\lambda\left[|n_i - n_k|\right] = \begin{cases} \rho_\lambda[0] = \sum_{n=0}^{\Lambda-1} h_\lambda^2[n], & i = k \\ 0, & i \neq k\ . \end{cases} \tag{4.37}$$

Hence, the solution to Equation 4.24 becomes

$$x_\phi[n_k] = \frac{1}{\rho_\lambda[0]}\nu_\phi[n_k] = \frac{1}{\rho_\lambda[0]} \sum_{n=n_k}^{n_k+\Lambda-1} e_0[n]h_\lambda[n - n_k]\ , \tag{4.38}$$

where n_k, $k = 1, 2, \ldots, K$, is defined by the RPE grids

$$n_k = \phi + (k-1)D,\ \phi = 0, 1, \ldots, D-1\ . \tag{4.39}$$

The optimum grid position on the basis of Equation 4.13 is now given by

$$p = \arg\max_\phi \sum_{k=1}^{K} x_\phi[n_k]\rho_\lambda[0]x_\phi[n_k] = \arg\max_\phi \sum_{k=1}^{K} x_\phi^2[n_k]\ , \tag{4.40}$$

and the optimum amplitudes by

$$x[n_k] = x_p[n_k]\ . \tag{4.41}$$

4.3.5 Full exploitation of the autocorrelation method

Further simplification of the above expressions for MPE and RPE can be achieved by recognising that $e_0[n]$ is the result of the convolution of $d[n]$ and $h_\lambda[n]$ (see Figure 4.1), since the filter $H(z) = 1/A(z/\gamma)$ and consequently also the FIR filter with impulse response h_λ, have zero initial state in the autocorrelation method. Hence,

$$e_0[n] = \sum_{j=0}^{N-1} d[j]h_\lambda[n - j]\ . \tag{4.42}$$

Let us now define a new sequence $y[i]$ in accordance with the sequence of Equation 4.35 as

$$y[i] = \frac{1}{\rho_\lambda[0]} \sum_{n=i}^{i+\Lambda-1} e_0[n]h_\lambda[n - i]\ . \tag{4.43}$$

Substitution of Equation 4.42 into this expression and rearranging the summations yields

$$y[i] = \frac{1}{\rho_\lambda[0]} \sum_{j=0}^{N-1} d[j] \sum_{n=i}^{i+\Lambda-1} h_\lambda[n-j] h_\lambda[n-i] = \frac{1}{\rho_\lambda[0]} \sum_{j=0}^{N-1} d[j] \rho_\lambda[i-j], \tag{4.44}$$

which can be interpreted as the convolution

$$y[n] = \frac{d * \rho_\lambda[n]}{\rho_\lambda[0]} \; . \tag{4.45}$$

As a remarkable result, the MPE coder using the autocorrelation method and radical truncation boils down to the system as shown in Figure 4.9. The prediction residual $d[n]$ is convolved with a time function $\rho_\lambda[n]/\rho_\lambda[0]$

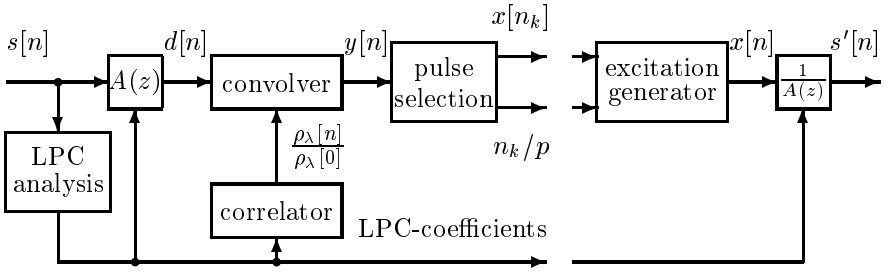

Figure 4.9: MPE and RPE coding scheme (left: encoding, right: decoding) resulting from autocorrelation analysis combined with radical truncation of the impulse response. In the RPE case, the pulse positions n_k are represented by the grid position p.

and the MPE pulses are selected from the resulting sequence $y[n]$. The convolver is different from a filter in the sense that the convolution is "block-wise" as a result from the autocorrelation method. In addition, the convolver is non-causal and it does not introduce any delay. The sequence $y[n]$ may be different from zero in the range $-(\Lambda-1) \leq n \leq N-1+(\Lambda-1)$, but the samples outside the range $0 \leq n \leq N-1$ are not used. In the sequential search, each pulse location is selected in the *allowed* region of n, within the range $0 \leq n \leq N-1$ and outside the regions of Equation 4.32, on the basis of Equations 4.35 and 4.43, by

$$n_k = \arg\max_n |y[n]| \tag{4.46}$$

and the associated pulse amplitude according to Equations 4.36 and 4.43, is given by

$$x[n_k] = y[n_k] \; . \tag{4.47}$$

In the case of RPE the pulse locations are always within the allowed region if $D \geq \Lambda$. The pulse amplitudes are, according to the Equations 4.38, 4.39 and 4.43, given by

$$x_\phi[n_k] = y[\phi + (k-1)D] \; , \tag{4.48}$$

and the optimum grid position p is, according to Equation 4.40, given by the sequence $x_\phi[n_k]$ having the highest energy.

It becomes clear by now that the autocorrelation method combined with radical truncation yields an RPE system which is less complex than the resulting MPE system. Although the differences are not very big, MPE still has the sequential search and the constraint of the allowed regions with the associated book-keeping. In addition, the pulse positions in MPE have to be coded in a bit-rate efficient way, before transmission. Some ways of accomplishing this are described in [11] and [14].

Next, we will explore what smart forms of a fixed system impulse response can bring in.

4.3.6 Fixed impulse response adjusted to speech

In this section, the perceptual and computational effects of making the filter $H(z)$ fixed according to the long-term average spectrum of speech are discussed.

The transfer function of this fixed filter will be denoted by $H_f(z) = 1/F(z)$ and it is defined by

$$H_f(z) = \frac{1}{F(z)} = \frac{1}{1 - P_f(z)} \; , \tag{4.49}$$

where $P_f(z)$ stands for the same fixed predictor as defined by Equation 3.18 in Section 3.2.2,

$$P_f(z) = \sum_{m=1}^{M} a_m z^{-m} \; . \tag{4.50}$$

That section deals with DPCM and predictors amongst which the optimum fixed predictor for speech signals according to Paez and Glisson [8]. The coefficients a_m for orders $M = 1, 2, 3$ are given in Table 3.1.

The performance of this filter for $M = 3$ in the spectral shaping of the coding error has been investigated in an experiment [11, 13, 18]. This experiment concerned two representative systems according to Figure 4.10, which use quantization noise as a substitute for the coding error. The systems use a quantizer Q on the weighted prediction resid-

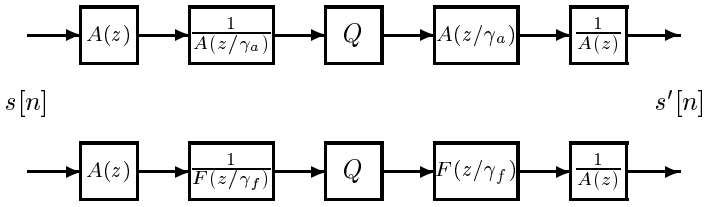

Figure 4.10: Systems used in the noise shaping experiment. Top: adaptive noise shaping. Bottom: fixed noise shaping.

ual. The unweighted residual is generated with the aid of a 12^{th} order inverse filter $A(z)$ the coefficients of which are refreshed every 10 ms and derived from 30 ms strongly overlapping segments, so that subsequent sets of coefficients change smoothly. The uniform 8-level forward adaptive APCM quantizer Q produces audible noise and it is supposed to generate an approximately white quantization noise spectrum. Before the quantizer output is fed to the synthesis filter $1/A(z)$, it is weighted by $A(z/\gamma_a)$ in an adaptive noise shaping version of the system and by $F(z/\gamma_f)$ in the fixed noise shaping system version. The weighting filters before the quantizers take care that the total transfer function from the input speech signal $s[n]$ to the output signal $s'[n]$ is always one, if the quantizer is left out of consideration.

In the adaptive version $\gamma_a = 0.8$ is used for obvious reasons and in the fixed version a range of values $\gamma_f = 1.0, 0.8, 0.6, 0.4, 0.2$ is used. The associated transfer functions of the fixed filter are shown in Figure 4.11. The unexpected result of informal listening tests with several experienced listeners was that the fixed filter versions with $\gamma_f = 0.6$ or 0.8 were preferred, and that the system with these parameters sounded at least as pleasant as the system with the adaptive filter. The filters with $\gamma_f = 0.6$ and 0.8 are characterised by a smooth curve with an attenuation difference between the frequencies 0 and 4000 Hz of 15 to 20 dB, respectively.

This makes the fixed filter a good candidate, but the computational

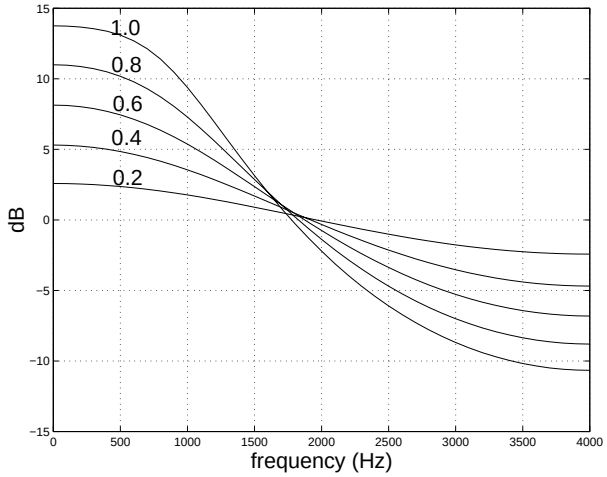

Figure 4.11: Transfer functions $|1/F(e^{j\theta}/\gamma_f)|$ of the fixed filter for γ_f =1.0, 0.8, 0.6, 0.4, and 0.2.

efficiency is limited to pre-calculation of the inverse matrices as discussed in Section 4.2.3 and the efficiency of the autocorrelation method combined with radical truncation cannot be achieved without further sophistication.

Although the impulse response of the fixed filter can also be subject to radical truncation, it is sufficient if the coefficients $\rho[j]$ of the matrix $\mathbf{R}$ of the autocorrelation method equal zero at $|j| = kD$, $k = 1, 2, \ldots, K-1$. In this case, a diagonal matrix is obtained as well and explicit solutions are readily obtained, again. Moreover, it will appear that this approach is also viable for $D = 3$, while for radical truncation the case of $\Lambda \leq 3$ has not been investigated, also because the expectations are that this case may be too radical and will not preserve the desired spectral properties. The strategy of tuning the fixed filter to the decimation factor is only applicable to RPE, of course. It will be elaborated in the next section.

4.3.7 Constructing explicit RPE-tuned forms

What is needed is a symmetrical time function $\rho_D[n]$, where the subscript indicates tuning to the decimation factor, that equals zero at the regular intervals $n = kD$, $k = \pm 1, \pm 2, \ldots, \pm(K-1)$ and the Fourier transform of which shows a smooth curve with an attenuation of 15 to 20 dB over the $0 \leq \theta \leq \pi$ frequency range. It is possible to construct

such a time function by the product of two functions,

$$\rho_D[n] = \beta[n]\frac{\sin n\pi/D}{n\pi/D}\ , \tag{4.51}$$

in which the sinc-function enforces the required zeros and $\beta[n]$ provides for the required attenuation properties. The sinc-function gives rise to a rectangular spectrum with edges at the frequencies $|\theta| = \pi/D$. This function is convolved with the Fourier transform $B(e^{j\theta})$ of $\beta[n]$ to yield the Fourier transform of $\rho_D[n]$. In order to obtain the required smooth result, a Butterworth-like frequency function $B(e^{j\theta})$ is defined as

$$B(e^{j\theta}) = \frac{1}{\sqrt{1 + \left(\frac{\theta}{\theta_c}\right)^{2m}}}\ , \tag{4.52}$$

where m stands for the order of the function and θ_c for the cut-off frequency. The values of θ_c and m are determined empirically to arrive at the desired results. To this end $B(e^{j\theta})$ is sampled as

$$B[i] = B(e^{j2\pi i/256}),\ i = 0:128 \tag{4.53}$$

and

$$B[256 - i] = B[i],\ i = 1:127 \tag{4.54}$$

to fit a 256-point (inverse) FFT giving $\beta[n]$. By trying several values for θ_c and m, multiplying each $\beta[n]$ by the sinc-function and observing the spectrum of the product, the results of Table 4.1 have been generated. The time sequence $\rho_4[n]/\rho_4[0]$ of Table 4.1 has been obtained by $\theta_c = 2\pi 800/8000$ and $m = 3$ and the sequence $\rho_3[n]/\rho_3[0]$ by $\theta_c = 2\pi 550/8000$ and $m = 2.7$. The obtained sequences have been *truncated* to the shown lengths. The truncated elements of the sequences concern function values which are well below 0.01. The amplitude spectra of the two truncated functions are given in Figure 4.12. The difference in attenuation at the frequencies 0 and 4000 Hz for the filter for $D = 4$ amounts to approximately 19 dB and for the filter for $D = 3$ to approximately 18 dB. These numbers are well within the desired range.

The use of such a convolver reduces the coding scheme of Figure 4.9 to the scheme of Figure 4.13. The sequence $y[n]$ is, in agreement with Equation 4.45, determined again by

$$y[n] = \frac{d * \rho_D[n]}{\rho_D[0]}\ , \tag{4.55}$$

$\|n\|$	$\rho_4[n]/\rho_4[0]$	$\rho_3[n]/\rho_3[0]$
0	1.00000	1.00000
1	0.70699	0.70079
2	0.28847	0.25079
3	0.05896	0.00000
4	0.00000	-0.04565
5	0.00724	-0.01636
6	0.01263	0.00000
7	0.00620	
8	0.00000	

Table 4.1: Symmetrical time functions $\rho_D[n]/\rho_D[0]$ for $D = 4$ and $D = 3$.

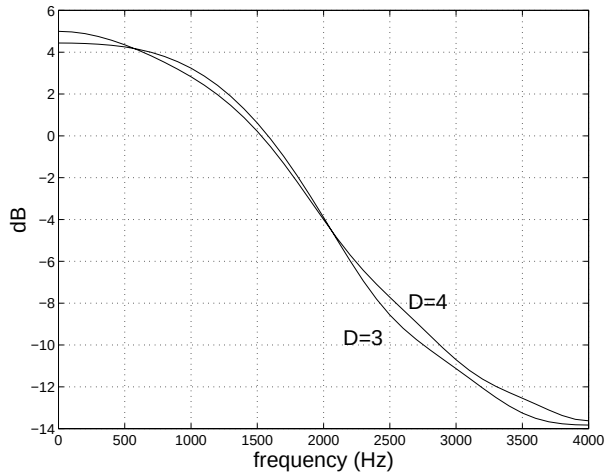

Figure 4.12: Frequency responses of $|1/F(e^{j\theta}/\gamma_f)|$, for a sampling rate of 8 kHz, of the RPE-tuned fixed filters for $D = 3$ and $D = 4$.

and the optimum grid position p is determined by selecting the D candidate sequences $x_\phi[n_k]$ according to Equation 4.48 and finding the candidate with the highest energy according to Equation 4.40. The pulse amplitudes are then given by $x[n_k] = x_p[n_k]$.

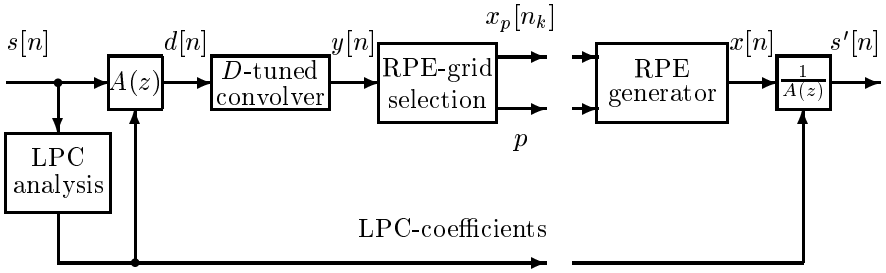

Figure 4.13: RPE coding scheme (left: encoding, right: decoding) resulting from autocorrelation analysis combined with decimation-(D-)tuned fixed system impulse response.

4.3.8 Some conclusions

On speech quality

In the literature it has been mentioned that quality loss due to the use of the autocorrelation method is not clearly detectable. This was supported by own experience.

It has been shown that, by radical truncation of the system impulse response to a duration of only 4 sampling periods, the perceptual error weighting function is not significantly affected.

It has been argued that the use of a fixed system filter according to the long-term average speech spectrum does not affect the perceptual noise shaping performance in a negative way.

In addition, it has been shown that tuning this fixed filter to an RPE decimation factor of 3 or 4 does not really change its noise shaping performance.

On computational efficiency

By the application of the autocorrelation method in combination with radical truncation to 4 samples, the analysis-by-synthesis system of Figure 3.19 reduces to the system of Figure 4.9, both for MPE and RPE. This system has a much simpler, RELP-like, architecture.

By applying the autocorrelation method in combination with a D-tuned fixed system filter, an even simpler architecture is obtained. This

system is shown in Figure 4.13. It is essentially based on RPE and its feasibility has been shown for decimation factors $D = 4$ and even $D = 3$.

It is only the autocorrelation function of the impulse response of the D-tuned fixed system filter that has to be known. By applying LP analysis to this autocorrelation function an IIR estimation of the filter itself and its impulse response can be obtained.

The resemblance of the latter system to the RELP architecture of Figure 3.15 with spectral folding is striking. See also the RELP system in [10]. The most important difference is that the LPF-design issue is inherently solved by the emerged convolver and that the straight-forward down-sampler is replaced by the more sophisticated RPE down-sampling including grid selection. Compare the sharp RELP filter of [10] (see also Section 3.4.3) to the filter of Figure 4.12 which allows substantial aliasing! It should be emphasised that this RPE system still has the property, as inherited from the analysis-by-synthesis architecture, to minimize the perceptual coding error.

The feasibility of a real-time hardware implementation of this system is proven by the fact that similar RELP systems already had been realized, as discussed in Section 3.4.3. Although these realizations needed more than one DSP in the early eighties, the situation in the mid-eighties was that implementation on the basis of a single DSP was a fact already [19].

On the basic approach

It should be noted that the methods of radically truncated and RPE-tuned fixed system impulse responses force explicit solutions *by construction*. Other reported strategies, described by Kroon in 1985 [14] and by Kroon, Sluijter and Deprettere in 1986 [18], are based on the observation that the matrices often show strong diagonal dominance, and that therefore $\mathbf{C}$ or $\mathbf{R}$ may be *approximated* by a diagonal matrix with the same diagonal. This is not necessarily always an adequate assumption.

4.4 Final architecture and design details

In the eventually standardized coder, which will be reviewed from Subsection 4.4.2 onwards, the explicit RPE approach with a D-tuned fixed system filter has been adopted, including pitch prediction. In the standard, pitch prediction is referred to as long-term prediction (LTP). LTP was introduced and implemented into the standard codec by Galand and

Rosso of IBM France [22]. Several embodiments of the RPE coder with pitch prediction, including the one in the standard, had already been covered by the earlier Philips RPE patent, though [23]. In the next subsection, we first introduce LTP into the explicit RPE system with the D-tuned fixed system filter.

4.4.1 Including LTP

The basic functionality of including LTP is shown in Figure 4.14, where the D-tuned fixed system filter is indicated by $H_D(z)$. Starting from the

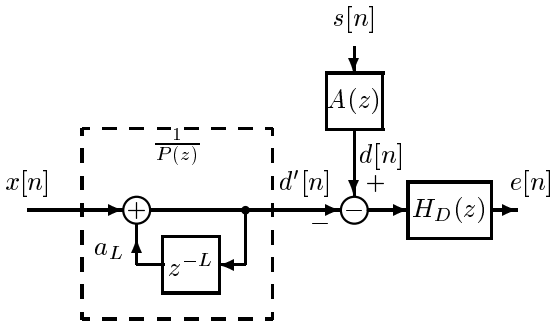

Figure 4.14: Analysis-by-synthesis scheme with pitch synthesis.

basic analysis-by-synthesis system configuration of Figure 4.1, a pitch synthesiser is included into the excitation path with a lag representing a pitch period of L samples and a prediction coefficient a_L (see Section 2.5.5). By proper adaptation of these parameters to the short-term prediction residual of the speech signal, the excitation signal $x[n]$ is freed from the task of generating pitch periodicity and it only has to approximate the noise-like residual of the cascade of a short-term predictor and a long-term predictor applied to the speech signal. In the sequel we will refer to this residual as the "LTP residual". The approximation to the short-term prediction residual $d[n]$, which itself will be referred to as "STP residual", is indicated by $d'[n]$. The benefits of LTP have been discussed in Section 2.5.5.

Figure 4.15 shows a circuit that is functionally equivalent to the circuit of Figure 4.14. If the signal path from $s[n]$ to $e[n]$ is observed, the excitation signal $x[n]$ is subsequently added to this path and subtracted from this path again, so that any excitation signal $x[n]$ results

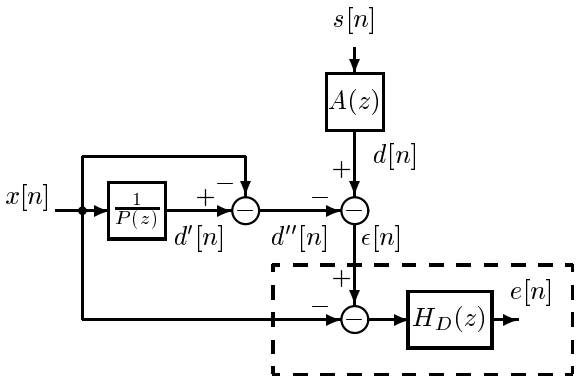

Figure 4.15: The same analysis-by-synthesis scheme subdivided into a part with pitch prediction and a part with a basic analysis-by-synthesis configuration according to Figure 4.1 (dashed box).

into the same error signal $e[n]$ in both circuits. The part of the diagram enclosed by the dashed box, however, is identical to the basic analysis-by-synthesis system configuration of Figure 4.1, but now with $\epsilon[n]$ as an input. Minimization of the energy of $e[n]$ over a subsegment has in the previous sections resulted in the explicit solution as shown in Figure 4.13. The parameters of the sequence $x[n]$ can be found by applying the D-tuned convolver to $\epsilon[n]$ and subsequent RPE-grid selection. Figure 4.16 shows the resulting architecture in which an RPE generator has been added to generate the sequence $x[n]$ from the resolved parameters, i.e. grid position p and pulse amplitudes $x_p[n_k]$, as an input to the LTP subsystem, also indicated by a dashed box. As a remarkable result, the analysis-by-synthesis loop has now turned into a DPCM-like loop with LTP in its feedback loop, but on a subsegmental basis since the explicit solution works on this basis. The signal $x[n]$ is an approximation of $\epsilon[n]$ which itself is an approximation of the LTP residual. In the Figure, $d''[n]$ is the (long-term) prediction from the reconstructed short-term prediction residual $d'[n]$.

Thus, an adequate decoder consists of an RPE generator, an LTP synthesiser with transfer function $1/P(z)$ giving the signal $d'[n]$ which excites a short-term synthesis filter with transfer function $1/A(z)$, resulting into the reconstructed speech signal $s'[n]$.

The LTP parameters L and a_L are determined, as one would expect,

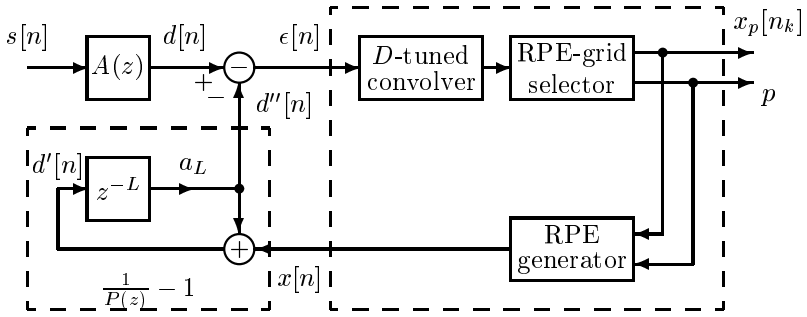

Figure 4.16: The same analysis-by-synthesis scheme with substitution of the explicit RPE solution using a D-tuned convolver, into the dashed box of the previous figure. This yields the right-hand dashed box. The left-hand dashed box contains the resulting LTP subsystem.

by minimising the energy of the signal representing the LTP prediction error, that is $\epsilon[n]$. This signal is given by

$$\epsilon[n] = d[n] - d''[n] = d[n] - a_L d'[n-L] \; . \tag{4.56}$$

Accordingly, a total square error E_ϵ over a subsegment can be defined by

$$E_\epsilon = \sum_{n=0}^{N-1} \epsilon^2[n] \; , \tag{4.57}$$

where N stands for the length of the subsegment. By putting the partial derivative of E_ϵ with respect to a_L to zero, we find the value of the prediction coefficient

$$a_L = \frac{\sum\limits_{n=0}^{N-1} d[n]d'[n-L]}{\sum\limits_{n=0}^{N-1} d'^2[n-L]} \; . \tag{4.58}$$

The total square error is then represented by

$$E_\epsilon = \sum_{n=0}^{N-1} d^2[n] - \frac{\left(\sum\limits_{n=0}^{N-1} d[n]d'[n-L]\right)^2}{\sum\limits_{n=0}^{N-1} d'^2[n-L]} \; , \tag{4.59}$$

so that the optimum lag L can be found by

$$L = \arg\max_i \frac{\left(\sum_{n=0}^{N-1} d[n]d'[n-i]\right)^2}{\sum_{n=0}^{N-1} d'^2[n-i]} \,. \tag{4.60}$$

So, the coding process first determines L and a_L on the basis of $d[n]$ and $d'[n-i]$. The latter values can be obtained from the contents of the delay line z^{-L}. The range of i over which is searched for the optimum pitch lag has a constraint, however. The values $d'[n-i]$ cannot be used if $n \geq i$. This case of $i \leq n$ can occur for values of i lower than the subsegment length N. In that forbidden case samples of d' would be needed with an argument of the current subsegment, but these are yet to be determined since we discuss the process to determine $\epsilon[n]$. So, recursion in the LTP subsystem within the current subsegment cannot be handled by this architecture. In the case of subsegments with a length of $N = 40$ samples, pitch period values below 40 have to be dealt with by forcing i to span multiple periods so that $i \geq N$. In Section 2.1 we have seen that we can expect pitch periods as short as 20 samples, indeed.

4.4.2 The standard codec

The block diagram of the standard encoder is shown in Figure 4.17 and the block diagram of the decoder is given in Figure 4.18. The codec is based on the system given in Figure 4.16. The sampling frequency is 8 kHz. All functions are implemented on the basis of 16-bit fixed-point arithmetic, in order to facilitate implementation in a digital signal processor. The encoder consists of three main functional parts, i.e. LPC, LTP and RPE. The decoder consists of the same main parts as the encoder, but in the reverse order. In the next subsections, the three parts LPC, LTP and RPE are discussed. As little as possible will be repeated which has already been described in the standard [26]. Emphasis will be put on the backgrounds and the underlying technologies, thus providing better understanding of the codec.

4.4.3 LPC

The speech input signal $s[n]$ of the encoder is subject to some preprocessing, consisting of a DC suppressing notch filter and a pre-emphasis

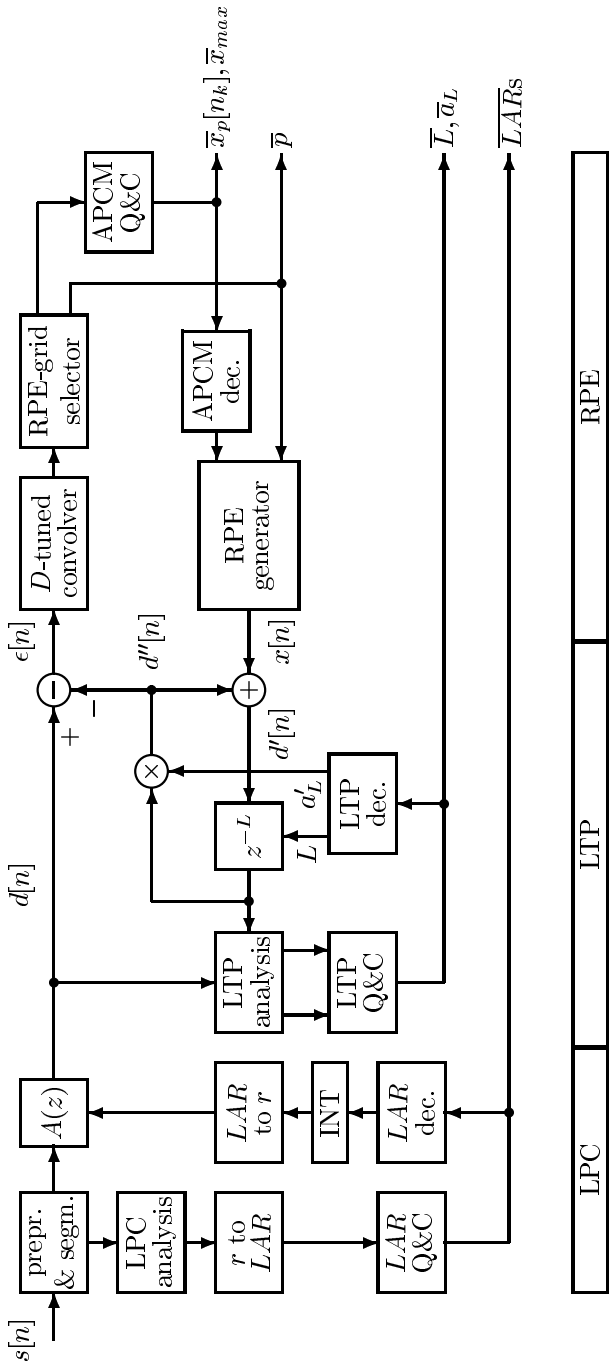

Figure 4.17: Block diagram of the standard encoder.

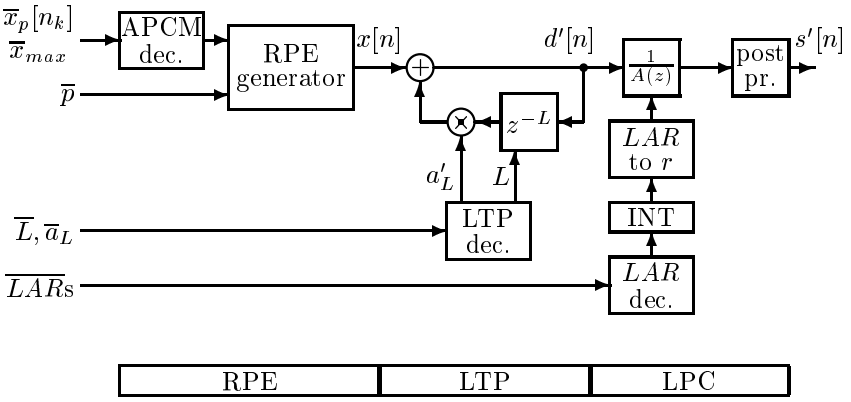

Figure 4.18: Block diagram of the standard decoder.

circuit. Subsequently, it is split up into segments of 20 ms for the LPC analysis. In the figure, these functions are indicated by the block "prepr.&segm.".

The pre-emphasis has a transfer function $1 - 0.86z^{-1}$, the absolute frequency response of which corresponds closely to the inverse of the first-order approximation to the long-term average speech amplitude spectrum according to Table 3.1. This will improve the quantization properties of the LPC coefficients as explained in Section 3.3.3. On the other hand, the properties of the STP residual $d[n]$ do not change because it still is the maximally whitened version of $s[n]$ given the constraints of the LPC procedure such as the order and accuracy, of course. In the decoder, the synthesis filter $1/A(z)$ and de-emphasis filter (indicated by the block "post pr.") together have a transfer function which is inverse with respect to the cascade of the pre-emphasis and $A(z)$, so that the noise shaping properties of the system are not changed by the application of the pre- and de-emphasis functions. An additional advantage of the (integrating) de-emphasis is that possible annoyingly audible spike-like effects due to transmission errors are smeared out in time.

The LPC analysis works with rectangularly windowed segments of pre-emphasised speech. This minimises the algorithmic delay as compared to the usual tapered, overlapping windows which require some look-ahead for each segment [20, 21]. Implementing such an overlapping scheme would increase the delay. The order of the LPC is 8. The LPC is based on the autocorrelation method and it uses the Schur recursion,

as explained in Section 2.5. In every segment eight reflection coefficients r are produced which are converted to LAR-parameters ("r to LAR" in Figure 4.17) using a piece-wise linear approximation to Equation 3.23 according to

$$LAR_m = \begin{cases} r_m \ , & |r_m| < 0.675, \\ \text{sign}(r_m)(2|r_m| - 0.675) \ , & 0.675 \leq |r_m| < 0.950, \\ \text{sign}(r_m)(8|r_m| - 6.375) \ , & 0.950 \leq |r_m| \leq 1.000. \end{cases} \quad (4.61)$$

This approximation avoids time-consuming logarithm and division computations in the fixed-point arithmetic. Subsequently the LARs are quantized and coded (LAR Q&C) into a total of 36 bits per set. The coded transmission parameters are denoted by $\overline{LAR}$s. Other parameter transmission codes are also denoted over-lined.

The quantization method of the LARs has been derived from the so-called "reference procedure" (RP), which was designed by the author in the early eighties and it was already used in the RELP coders decribed in [9] and [10]. The name RP stems from the fact that Kroon used it in his thesis [14] as a reference in a comparison of several quantization methods for the LPC coefficients. These comparisons showed the superior performance of this method, although the required bit rate is relatively high. The RP was a low complexity compromise between the US and MD methods discussed in Section 3.3.3. Just as in the US method, the RP method uses simple uniform quantizers on the LARs, but it also makes use of some statistics. The RP method uses the empirically determined perceptually relevant outer boundaries of the LAR probability density functions. These functions have been estimated from histograms from almost 18000, Hamming windowed not pre-emphasised, speech segments. The speech material was produced by two native Dutch speakers (male and female) and two native English speakers (also male and female). The minimum deviation method (see Section 3.3.3), on the contrary, utilises the whole probability density function according to Equation 3.32. Other statistical information used by the RP method is the fact that the spectral sensitivity decreases slightly with increasing index of the LAR, as later confirmed by Kroon in Section 6.4 of his thesis. This is exploited in the RP method by doubling the quantization steps from the 5^{th} LAR onwards. The RP method is summarised in Table 4.2 which shows the quantization steps, the applied outer boundaries (min., max.) and the bit allocations. The total number of bits for the 12^{th} order RP quantizer adds up to 52 bits per set. An 8^{th} order RP quantizer uses 38 bits per set. This procedure has been modified [24] to fit the new

LAR	min.	max.	step	bits
1-2	-1.60	1.55	0.05	6
3	-1.00	0.55	0.05	5
4	-0.55	1.00	0.05	5
5-10	-0.70	0.80	0.10	4
11-12	-0.30	0.40	0.10	3

Table 4.2: LAR quantization according to RP.

LAR	min.	max.	step	bits
1-2	-1.60	1.550	0.050	6
3	-1.00	0.550	0.050	5
4	-0.55	1.000	0.050	5
5	-0.60	0.500	0.073	4
6	-0.30	0.700	0.067	4
7	-0.40	0.440	0.120	3
8	-0.20	0.593	0.113	3

Table 4.3: LAR quantization in the GSM coder.

conditions of pre-emphasised and rectangularly windowed speech. The result, as shown in Table 4.3, yields a bit rate of only 36 bit per set while the first four LAR quantizers are still the same as in the RP case.

The $\overline{LAR}$s are sent to the decoder, and they are also locally decoded (LAR dec.), linearly interpolated (INT) and converted to reflection coefficients again (LAR to r).

The interpolation, which has to prevent transient effects due to updating the reflection coefficients in the inverse filter $A(z)$, provides for a smooth fade-out/fade-in of the coefficients of the previous segment and those of the current segment. In the early eighties it was common knowledge that interpolation could better be done on the LARs than on the reflection coefficients. This has later been confirmed in 1989 in a paper by Atal, Cox and Kroon [15], amongst others. The LAR-interpolation method used in the GSM codec deviates from the conventional linear interpolation over 5-ms subsegments. It is concentrated in the first 5 ms of a segment in order to put forward the time instant from which the actual set of LARs is used, whereas conventional interpolation methods use the coefficients of the previous segment until far into the current

segment (often into the third subsegment). The new approach gives a better match between the speech segment and the LPC coefficients. During the remaining 15 ms of a segment the coefficients of the current segment are used.

Recovering the reflection coefficients from the decoded LARs is done by the inverse process of Equation 4.61. This gives

$$r_m = \begin{cases} LAR_m\,, & |LAR_m| < 0.675, \\ \text{sign}(LAR_m)(0.337500 + |LAR_m|/2), & 0.675 \le |LAR_m| < 1.225, \\ \text{sign}(LAR_m)(0.796875 + |LAR_m|/8), & 1.225 \le |LAR_m| \le 1.625. \end{cases} \tag{4.62}$$

In the decoder, the received $\overline{LAR}$s are subject to the same decoding and interpolation procedure and used by the synthesis filter $1/A(z)$, which reproduces the pre-emphasised speech signal from the reconstructed STP residual $d'[n]$ which is delivered by the LTP part. Hence, $A(z)$ in the encoder and $1/A(z)$ in the decoder are guaranteed to be reciprocal, in the absence of transmission errors of course, because they always operate with the same reflection coefficients. This enhances the transparency of the codec [10]. After post processing (post pr.), consisting of a de-emphasis which compensates for the pre-emphasis, the reconstructed speech signal $s'[n]$ is recovered.

The inverse filter $A(z)$ is realized as a lattice filter according to Figure 2.28. The synthesis filter $1/A(z)$ is realized as the lattice filter of Figure 2.27. This prevents the need for forward and backward recursions between reflection coefficients and the direct-form a-parameters. Thanks to the use of the autocorrelation method, the Schur recursion and lattice filters, no a-parameters occur in the whole codec and only well-bounded reflection coefficients and LARs occur, thus facilitating fixed-point arithmetics. In addition, because the autocorrelation method is used, each section in the lattice filter $A(z)$ will reduce the signal level since it will add some prediction gain according to Equation 2.75, or, in the worst case, will not increase the signal level ($r_m = 0$). These well defined signal boundaries also strongly support the fixed-point implementation. Because the reconstructed speech signal is still the (perceptually weighted) mean square approximation of the source signal, the same is true for the synthesis filter $1/A(z)$. In direct-form implementations, the internal signal levels are not so well defined.

4.4.4 LTP

The LTP part works on 5-ms subsegments, both in the encoder and the decoder. The analysis algorithm of the standard is based on Equation 4.58 and

$$L = \arg\max_i \sum_{n=0}^{N-1} d[n]d'[n-i] \;, \tag{4.63}$$

respectively. The latter expression, however, is not in agreement with Equation 4.60 which represents the covariance method according to Section 2.5.5. Nor does it comply with the autocorrelation method, as described in the same section. The formulation of the autocorrelation method in this case is problematic, anyhow, since data from long ago may be contained in the memory of the delay line z^{-L}. Nevertheless, the results using Equation 4.63 do not deviate very much from those using Equation 4.60, because in most practical cases, large variations of the energy term represented by the denominator of Equation 4.60 often occur relatively slowly.

The range of i is limited by the subsegment length, i.e. $L_{min} = 40$, as discussed in Section 4.4.1. The upper limit is set to $L_{max} = 120$, because exclusion of longer periods from prediction saves considerable computational effort. On the other hand, RPE can handle well periods of this length or longer, on its own, and the pitch predictor does not clearly contribute anymore to the speech quality under these conditions [9]. Therefore, the standard codec maintains a memory of 120 previous samples of $d'[n]$.

Quantization of a_L is realized by a non-uniform quantizer. It is guaranteed that $a_L \leq 1$ and consequently the pitch synthesis filter is guaranteed to be stable.

Hence, the quantization and coding (LTP Q&C) results in a 7 bits code for $\overline{L}$ and a 2-bit code for $\overline{a}_L$, which are transmitted to the decoder. Locally, they are decoded (LTP dec.) to be used in the long-term predictor to generate $d''[n]$ out of the memory contents of the delay line z^{-L} together with the reconstructed prediction parameter a'_L. By subtraction of this signal from the STP residual $d[n]$, the estimated LTP residual $\epsilon[n]$ is obtained. Processing of the next subsegment will be prepared by computing $d'[n]$ out of $d''[n]$ and $x[n]$, as soon as the latter becomes available, and shifting $d'[n]$ into the delay line. In the decoder the received LTP parameters are decoded in the same way as in the encoder and the LTP synthesiser generates $d'[n]$ from the excitation signal $x[n]$, which is delivered by the RPE part.

4.4.5 RPE

The RPE part in the encoder contains the D-tuned convolver, the RPE-grid selector and the RPE generator which generates the excitation signal $x[n]$ according to Figure 4.16. The applied decimation factor is $D = 3$. The convolver coefficients are given in Table 4.1. Because $D = 3$ does not fit to the subsegment length $N = 40$, there are four grids according to

$$n_k = \phi + (k-1)D, \quad k = 1, 2, \ldots, 13, \quad \phi = 0, 1, 2, 3. \tag{4.64}$$

The grid position p is selected out of these four possible grid positions using Equation 4.40 and $\overline{p}$ is represented by a 2-bit code. Thirteen pulse amplitudes $x_p[n_k]$, $k = 1, 2, \ldots, 13$, associated with the grid position p, are encoded on the basis of forward block-adaptive APCM (APCM Q&C) using an 8-level uniform mid-rise quantizer, see Section 3.1, resulting in a 3-bit code for each $\overline{x}_p[n_k]$. The block maximum x_{max} is determined by

$$x_{\max} = \max_k |x_p[n_k]| \,, \tag{4.65}$$

and encoded logarithmically into a 6 bits code $\overline{x}_{max}$. The quantizer inputs are normalised values $x'_p[n_k]$ given by

$$x'_p[n_k] = x_p[n_k]/x'_{\max} \,, \tag{4.66}$$

where x'_{max} is the decoded value of the block maximum. In this way, both in the encoder and in the decoder the same decoded RPE pulses are obtained. In the encoder a local decoder (APCM dec.) delivers the reconstructed pulse amplitudes to the RPE generator. In the remote decoder the same pulse amplitudes are applied to the RPE generator to recover the excitation signal $x[n]$.

4.5 Selection process and performance

Undeniably, essential forerunners of the RPE-LTP coder are the RELP coders discussed in Section 3.4.2 using a pitch predictor to solve the tonal noise problem [9, 10]. In the latter reference, these coders were called SRELP (second-residual excited linear prediction) coders. Table 4.4 shows some attributes of these two coders (No. 1 and No. 2 in the table), and of other coders which were involved in the selection process. The papers No. 8 and No. 11 are test reports. The SRELP codec No.

No.	reference	system	segment update (ms)	bit rate kbit/s	codec complexity
1	[9], 1984	SRELP	20	9.6	–
2	[10], 1985	SRELP	20	9.6	10 μPD7720
3	[16], 1985	RPE	16.5	16	5 μPD7720
	[17], 1985				
4	[19], 1986	,,	,,	,,	0.5 PCB5010
5	[20], 1986	,,	,,	,,	3 μPD7720
6A	[21], 1987	RPE	19.5	14.77	3 μPD7720 (1.5 MOPs/s)
6B	[21], 1987	RPE-LTP	19.5	13.38	4 μPD7720 (2.2 MOPs/s)
7	[22], 1988	MPE-LTP	20	13.2	82% of a proprietary processor
8	[24], 1988	RPE	19.5	14.77	1.5 MOPs/s
9	[25], 1988	RPE-LTP	20	13.0	0.6 PCB5010 0.4 DSP16 0.45 TMS320C25
10	[27], 1989	RPE-LTP	,,	,,	,,
11	[28], 1989	RPE-LTP	,,	,,	,,

Table 4.4: Predecessors and the final version of the GSM standard codec (No. 9 and 10).

2 had an architecture very much like the standard codec of Figure 4.17, albeit that it was not yet based on RPE. It already used a segment update rate of 20 ms, the Schur algorithm and the lattice filters, and the same piece-wise linear approximation of the LARs, LTP, the APCM coder for the decimated baseband and the logarithmically encoded block maximum, as in the GSM standard. The decimation factor was $D = 4$. The reported complexity was 10 NEC μPD7720 processors.

The first RPE coder (No. 3), did not need a pitch predictor because the speech quality was good enough thanks to the relatively high bit rate of 16 kbit/s. The applied D-tuned explicit-by-construction form of RPE was disclosed for the first time. The complexity was only 5 NEC μPD7720 processors. A newly developed signal processor PCB5010 by Philips, described in paper No. 4, was used to implement this RPE coder and it was reported that only one processor was required for a duplex codec. No. 5 represents another implementation of the same codec on

the basis of the NEC processors. The complexity could be reduced to only 3 μPD7720 processors by using additional memory. The paper also mentions a codec delay of 32 ms.

In the fall of 1986 quality tests were performed by the CEPT on six codecs, i.e. a sub-band coder from Italy (CSELT) at 15 kbit/s, a sub-band coder from Norway (ELAB, Trondheim) at 15 kbit/s, a sub-band coder from Sweden (Ellemtel Utveckling AB) at 13 kbit/s, a sub-band coder from UK (British Telecom) at 15 kbit/s, an MPE-LTP coder from France (IBM, La Gaude) at 13.2 kbit/s and the RPE-LPC coder from Germany (Philips Kommunikations Industrie AG, Nürnberg) at 14.77 kbit/s, as described in No. 8. The latter coder is listed as No. 6A in the table. All coders had a gross bit rate of 16 kbit/s including some error protection. The only difference between coder No. 5 and No. 6A is the segment update rate, and their complexities are the same, as indicated in the table. The complexity is next to the number of required μPD7720 processors also expressed in an alternative dimension of mega operations per second (MOPs/s), for the first time. The results of these tests, reported in No. 8, were clearly in favour of the pulse-excited coders. Another outcome of the tests was that the codecs were far superior to an analog FM voice channel, which was also included in the tests. The average score of the RPE-LPC coder was MOS=3.54 and the MPE-LTP coder (No. 7) scored MOS=3.27. The codec delay of RPE-LPC was reported to be < 40 ms. The average score included various speech input and listening levels, additive environmental noise, various representative (for the mobile radio channel) transmission error patterns, and several tandeming schemes representative for the application environment. The tandeming schemes included the ITU G.711 (64 kbit/s PCM) and G.721 (32 kbit/s ADPCM) representing interfacing to the PSTN and also the codec itself, representing mobile to mobile calls. See for further details paper No. 8. The complexity of the MPE-LTP was estimated at 4.9 MOPs/s, while the RPE-LPC codec only required 1.5 MOPs/s (see paper No. 9).

After the tests, the idea was born to combine the two pulse excited codecs to come to an even better solution. The RPE-LPC codec was combined with LTP and the resulting codec, as proposed in No. 6B, had a reduced bit rate of 13.38 kbit/s and it was expected that the speech quality could at least be maintained. This codec was called RPE-LTP and the price to be paid was an increase in complexity of about 0.7 MOPs/s.

The final version of this RPE-LTP coder, reported in No. 9 and No. 10, incorporated some modifications as compared to the original tested RPE-LPC version. Apart from the extension with LTP, it concerned some design details, i.e. using a fourth grid position with decimation three and using pre- and de-emphasis, a rectangular window in the LPC analysis, a reduced order (to eight) of the LPC and the modification of the quantization tables of the 4^{th} until the 8^{th} LAR, thus reducing the bit rate of the LARs to 36 bit per set, and a modified LAR-interpolation method. As a result, the bit rate was reduced again to 13.0 kbit/s and the codec delay was reduced to 28 ms. The gross bit rate including error protection had become 22.8 kbit/s. Several verification implementations have been made, including implementations on the Philips processor PCB5010 [19], the proprietary IBM processor DSP16 [29,30] and the well-known processor TMS320C25 of Texas Instruments. The required computational effort for the codec is given in the table, for each of these processors.

A final verification test was carried out, as reported in paper No. 11. The results confirmed that the average speech quality exceeded the original performance, so MOS > 3.54 and the codec delay was reported to amount to "approximately 25 ms".

4.6 Summary

The design issues of the full rate GSM speech codec, playing a role during the mid 1980s, have been highlighted. The codec has been built on a pulse-excited analysis-by-synthesis framework and the major challenge was to reduce the computational complexity of this approach while keeping its speech quality. Both MPE (multi-pulse excitation) and RPE (regular-pulse excitation) options have been considered until it became evident that RPE was the preferred choice.

The most complex part in the RPE analysis-by-synthesis framework was the solution of a set of (13×13) equations, every 5 ms. Several approaches to complexity reduction have been discussed, including the autocorrelation method for solving the set of equations and manipulations of the system impulse response in order to force explicit forms, thus avoiding the solution of the set of equations. These manipulations concerned truncation, radical truncation, presetting and fixing, and adjusting the fixed impulse response to the long-term average speech spectrum. The perceptual consequences of all these measures have been closely

monitored. The final solution was the combination of the autocorrelation method with a construction of a fixed system impulse response adjusted to the long-term spectrum of speech and also tuned to the RPE decimation factor. This combined explicit by construction autocorrelation approach has resulted in a robust, low-complexity RELP-like architecture while the speech quality has been kept virtually equal to the original quality of the full-fledged analysis-by-synthesis framework. Including LTP (long-term prediction) completed the final LPC-LTP-RPE codec architecture as it still is today, at a total bit rate of 13 kbit/s. Finally, the standardization process and performance tests have been recapitulated.

References

1. P. Kroon, E.F. Deprettere, R.J. Sluijter, Regular-pulse excitation: a novel approach to effective and efficient multi-pulse coding of speech, *IEEE Transactions on Acoustics Speech and Signal Processing*, October 1986, pp. 1054–1063.
2. H.J. Kotmans, R.J. Sluijter, T.A.C.M. Claasen, An hierarchical architecture for real-time digital signal processing hardware Part I: General considerations, *Philips Nat.Lab. Technical Note* No.70/82, April 1982.
3. R.J. Sluijter, H.J. Kotmans, T.A.C.M. Claasen, An hierarchical architecture for real-time digital signal processing hardware Part II: A micro array processor proposal, *Philips Nat.Lab. Technical Note* No.71/82, May 1982.
4. F.J.A. van Wijk, J.L. van Meerbergen, F.P.J.M. Welten, R.J. Sluijter, Integrated and programmable processor for word-wise digital signal processing, *USA patent* 4,689,738, (priority date) December 27, 1983. (granted August 25, 1987).
5. G.P. Nayyar, Multi-pulse Excitation source for speech synthesis by linear prediction, *Institute for Perception Research Report* no. 439, Eindhoven, April 1983.
6. M. Berouti, H. Garten, P. Kabal, P. Mermelstein, Efficient computation and encoding of the multi-pulse excitation for LPC, *IEEE International Conference on Acoustics Speech and Signal Processing*, San Diego, March 1984, pp. 10.1.1–10.1.4.
7. G.A. Senensieb, A.J. Milbourn, A.H. Lloyd and I.M. Warrington, A non-iterative algorithm for obtaining multi-pulse excitation for linear-predictive speech coders, *IEEE International Conference on Acoustics Speech and Signal Processing*, San Diego, March 1984, pp. 10.2.1–10.2.4.
8. M.D. Paez and T.H. Glisson, Minimum Mean-Squared-Error Quantization in Speech PCM and DPCM Systems, *IEEE Transactions on Communications*, Volume COM-20, April 1972, pp. 225–230.
9. R.J. Sluijter, G.J. Bosscha and H.M.P.T. Schmitz, A 9.6 Kbit/s Speech Coder for Mobile Radio Applications, *IEEE International Conference on Communications (ICC)*, Amsterdam, May 1984, pp. 1159–1162. (P. Dewilde and C.A. May (Eds.), *Links for the Future*, IEEE/Elsevier Science Publishers B.V., North Holland, 1984.)

10. P. Vary, R.J. Sluijter, Speech processing in the mobile radio terminal, *Nordic Seminar on Digital Land Mobile Radio Communication*, Espoo, Finland, February 1985, pp. 67–76.
11. R.J. Sluijter, Coding the prediction residual of speech, *Philips Nat.Lab. Technical Note* No.10/85, March 1985.
12. P. Kroon, E.F.A. Deprettere, and R.J. Sluijter, Multi-pulse excitation Linear-predictive speech coder, *European Patent* 0 195 487 B1, (priority date) March 22, 1985. (application published September 24, 1986, granted June 7, 1989) *USA patent* 4,932,061 (granted June 5, 1990).
13. E.F. Deprettere and P. Kroon, Regular excitation reduction for effective and efficient LP-coding of speech, *IEEE International Conference on Acoustics Speech and Signal Processing*, Tampa, March 26, 1985, pp. 965–968.
14. P. Kroon, Time-Domain Coding of (near) Toll Quality Speech at Bit Rates Below 16 kb/s, *Technical University Delft Ph.D. Thesis*, May 21, 1985.
15. B.S. Atal, R.V. Cox and P. Kroon, Spectral quantization and interpolation for CELP coders, *IEEE International Conference on Acoustics Speech and Signal Processing*, Glasgow, May 1989, pp. 69–72.
16. R.J. Sluijter, Multi-pulse and Regular-pulse Speech Coders in RELP-like forms, *Eurasip/Cost Project 207 Workshop on Medium Rate Speech Coding*, Hersbruck, Germany, September 25, 1985 (Abstract).
17. P. Vary, R. Sluijter, Sprachcodierung für mobile Telefonsysteme, *NTG Fachberichte Nr. 90: Bewegliche Funkdienste*, München, November 25, 1985, pp. 172–177.
18. P. Kroon, R.J. Sluijter and E.F. Deprettere, A low complexity regular pulse coding scheme with a reduced Transmission delay, *IEEE International Conference on Acoustics Speech and Signal Processing*, Tokyo, April 1986, pp. 3083–3086.
19. K. Hellwig, P.Vary, P. Anders, J.v. Meerbergen, R. Sluijter, F.v.Wijk, Architectural user aspects of the single chip digital signal processor PCB5010, *Proceedings EUSIPCO-86*, The Hague, The Netherlands, September 1986, pp. 1239–1242.
20. K. Hellwig, R. Hofmann, R.J. Sluijter, P. Vary, MATS-D speech codec: regular-pulse excitation LPC, *Second Nordic Seminar on*

Digital Land Mobile Radio Communication, Stockholm, Sweden, October 1986, pp. 257–261.

21. P. Vary, R.J. Sluijter, C. Galand, M. Rosso, RPE-LPC codec- the candidate for the GSM Radio Communication System, *International Conference on Digital Land Mobile Radio Communication*, Venice, Italy, June 1987, pp. 507–516.
22. C. Galand, M. Rosso, Ph. Elie and E. Lançon, MPE/LTP speech coder for mobile radio application, *Speech Communication*, Volume 7, No.2, July 1988, pp. 167–178.
23. R.J. Sluijter, Clarification of the Unique Relation between the European Patent EP 0 195 487 B1 and the GSM Recommendation 06.10 on Full-Rate Speech Coding, *Philips Research Memo*, RWR-537-RS-93085-rs, October 5, 1993.
24. J.E. Natvig, Evaluation of Six Medium Bit-Rate Coders for the Pan-European Digital Mobile Radio System, *IEEE Journal on Selected Areas in Communications*, Volume 6, No.2, February 1988, pp. 324–331.
25. P. Vary, K. Hellwig, R. Hofmann, R. Sluijter, C. Galand, M. Rosso, Speech codec for the European mobile radio system, *IEEE International Conference on Acoustics Speech and Signal Processing*, New York, April 1988, pp. 227–230.
26. GSM full rate speech transcoding, *ETSI/GSM Recommendation* 06.10, September 19, 1988.
27. K. Hellwig, P. Vary, D. Massaloux, J.P. Petit, C. Galand, M. Rosso, Speech Codec for the European Mobile Radio System, *IEEE Globecom Conference Record*, Dallas, November 1989, pp. 1065–1069.
28. A. Coleman, N. Gleiss, J. Sotschek, P. Usai and H. Scheuermann, Subjective performance evaluation of the RPE-LTP codec for the Pan-European cellular digital mobile radio system, *IEEE Globecom Conference Record*, Dallas, November 1989, pp. 1075–1079.
29. C. Galand, C. Couturier, G. Platel and R. Vermot-Gauchy, Voice-excited predictive coder (VEPC) implementation on a high-performance signal processor, *IBM Journal of Research and Development*, Volume 29, No.2, March 1985, pp. 147–157.
30. J.P. Beraud, Signal processor chip implementation, *IBM Journal of Research and Development*, Volume 29, No.2, March 1985, pp. 140–146.

Chapter 5

Epilogue: later developments

This final chapter briefly reviews the relevant developments in speech coding technology after the mid eighties. Especially in the nineties, many standards have been set for mobile telephony as well as for wired-network applications. Most of these standards are included in this chapter as state-of-the-art milestones. The algorithms according to these coders are not explained but their main features and innovative elements are discussed and key literature is listed. The chapter is concluded by an outlook on coding technology.

5.1 Setting the scene

The state of the art in 1988 is reported in an issue of the IEEE Journal on Selected Areas in Communications on voice coding for communications [1]. With regard to the successful narrow-band analysis-by-synthesis speech coding technology, the state of the art in the late eighties is very well reflected by the two papers [2] and [3]. In [2], presented in 1986, Trancoso and Atal report on their search for efficient algorithms for CELP coding (recall Section 3.5.2). One of their conclusions is that still several state-of-the-art DSPs are required to realize a real-time CELP codec. In the paper of Kroon and Deprettere from 1988 [3], multi-pulse and regular-pulse excitation are compared to code excitation with respect to quality performance as well as to complexity. Bit rates range from 16 down to 4.8 kbit/s. The observation is made that for similar quality performance the pulse-excited approaches yield faster algorithms

which can be implemented on a single state-of-the-art DSP while the codebook approach is more demanding, despite the use of *sparse codebooks*. In a sparse codebook the majority of the samples in an excitation vector are zero. They were constructed from severely center-clipped zero-mean Gaussian noise sequences.

The state of the art in wide-band speech coding technology in 1988 is reflected by the CCITT standard G.722 [5, 4]. It concerns a sub-band approach with *two quadrature mirror half-bands*, one conveying the lower-band signal and the other one conveying the remaining upper-band signal. In each band the signal is down-sampled by a factor of 2 to 8 kHz and independently coded with its own ADPCM coder. The total bit rate of this SB-ADPCM coder can be adjusted to 64, 56 or 48 kbit/s.

After 1988, the year of the approval and publication of the full-rate GSM standard, many standards have been set for mobile telephony as well as for wired-network applications. Most of them, without claiming completeness, are listed in Table 5.1. The table shows 25 standards, 14 of which (bold-faced print) have been designed for mobile telephony. Only 6 out of those 25 carry a speech-coding-type name or acronym not derived from CELP. In the early 1990s it had become common practice to use the word CELP for all analysis-by-synthesis coders. Variants to this acronym, such as VSELP and ACELP, almost always refer to the particular way of excitation applied, which is not always the codebook excitation which is referred to in the CELP-acronym. As will appear, the ACELP coder, for instance, shows more relation to a pulse-excited coder than to a codebook-excited coder. Nevertheless, we will comply with this inconsistent "everything-is-CELP" convention. A separate section, the next one, is entirely devoted to the development of this class of coders.

The other 6 types of coders concern the narrow-band ADPCM coder according to the ITU standard G.726 which has already appeared in Section 1.4 and has been explained in detail in Section 3.2.2, the above mentioned wide-band SB-ADPCM coder, 2 sinusoidal vocoders, 1 LPC vocoder called MELP (which is *not* a CELP variant) and a wide-band transform coder (TC). The vocoders will be discussed in Section 5.3. The TC is addressed in Section 5.4 on non-speech-specific coding.

year	name	institute	type	bit rate (kbit/s)
1988	G.722 (WB)	ITU-T	SB-ADPCM	48, 56, 64
1989	**IS-54**	TIA	VSELP	7.95
1990	G.726	ITU-T	ADPCM	16, 24, 32, 40
1990	**PDC Full Rate**	RCR	VSELP	6.7
1990	**IMBE**	Inmarsat	Sin. Vocoder	4.15
1991	FS1016	USA-DoD	CELP	4.8
1991	G.728	ITU-T	LD-CELP	16
1993	**PDC Half Rate**	RCR	PSI-CELP	3.45
1993	**IS-96**	TIA	QCELP	0.8, 4.2, 8.5
1994	**GSM-HR**	ETSI	VSELP	5.6
1995	G.729	ITU-T	CS-ACELP	8
1995	G.723.1	ITU-T	A/MP-CELP	5.3, 6.4
1995	**IS-127 (EVRC)**	TIA	RCELP	0.8, 4.2, 8.5
1995	**GSM-EFR**	ETSI	ACELP	12.2
1996	**IS-136-EFR**	TIA	ACELP	12.2
1996	FIPS Pub. 137	USA-DoD	MELP	2.4
1997	**IS-641**	TIA	ACELP	7.4
1999	G.722.1 (WB)	ITU-T	TC	24, 32
1999	**GSM-AMR**	ETSI	ACELP	4.75 - 12.2
1999	HVXC	MPEG-4	Sin. Vocoder	2 - 4
1999	CELP	MPEG-4	MPE-CELP	4 - 12
1999	CELP (WB)	MPEG-4	MPE/RPE-CELP	11 - 24
2001	**AMR-WB**	3GPP	ACELP	6.6 - 23.85
2001	**SMV**	3GPP2	RCELP	0.8, 2, 4, 8.55
2003	**VMR-WB**	3GPP2	ACELP	1.0 - 13.3

Table 5.1: Standards set after the GSM full-rate standard. The bold-faced standards have been created for mobile telephony, the others for wired-network applications.

5.2 Developments in CELP coding

In 1989 the VSELP, which stands for vector-sum excited linear prediction, is launched as the speech coder for the IS-54 digital cellular telephony radio system of the TIA [6]. The bit rate is 7.95 kbit/s. The VSELP algorithm, which is basically the CELP algorithm with a new excitation scheme, contained several other earlier proposed innovations.

Firstly, the concept of the *adaptive codebook* was applied. The delay line of the pitch prediction synthesiser contains past excitation samples of the short-term synthesis filter $1/A(z)$ and can be interpreted as a codebook with contents that has adapted itself to the recent excitation signal of $1/A(z)$. The LPC parameters of $1/A(z)$ are determined once every 20 ms and the coder works with 5-ms subsegments. Most importantly, the functionality of the adaptive codebook is associated with a provision that allows the lag L to take values less than the length of a subsegment (recall the last paragraph of Section 4.4.1). In the analysis algorithm of some CELP implementations the subsegment is completed in that case by the original prediction residual but in most implementations, including VSELP, the selected period of L samples from the delay line is repeated until the subsegment is completed. Thus, recursions within the subsegment under analysis are avoided and simple linear relations resulting from least-square-error criteria stay valid without any constraint on the range of L. This measure contributes to the speech quality.

Secondly, *sequential estimation*, as in suboptimum multi-pulse excitation (see Section 3.5.1 and 4.1.1) is applied. This means that in the analysis algorithm for a subsegment, first the contribution of the system memory contents to the perceptually-weighted coding error is determined and processed under the condition of zero excitation signal of $1/A(z)$. Subsequently, the optimum reduction by the adaptive codebook of the remainder of the coding error is determined under the condition of zero fixed-codebook excitation and processed to obtain an updated remainder of the coding error. One of the parameters in this process is the lag L which is varied so as to minimize the current (perceptually weighted) coding error. This way of analysis, which is applied in VSELP, is often referred to as *"closed-loop" pitch estimation.* In "open-loop" pitch estimation, on the contrary, the pitch period is directly estimated from the speech signal or its prediction residual and assigned to L, without regarding its effect on the coding error. As a last step in the sequential approach, the contribution of the fixed code-

book is determined by selecting the excitation vector which minimises the remaining coding error. The sequential approach is a great help in complexity reduction.

Thirdly, the *vector-sum structure* of the fixed codebook results in another reduction of the complexity, thus enabling real-time implementation of the VSELP on a single DSP [7]. The vector-sum structure is defined by 2^7 different excitation vectors formed out of linear combinations of only 7 basis vectors with weighting factors +1 or -1. This gives a 7-bit code. The advantage is that for a subsegment the contribution of each basis vector to the coding error can be determined in advance and that the selection process essentially reduces to finding the optimum linear combination of these contributions. This reduces the filtering of 128 excitation vectors, which constitutes the major part of the processing of the CELP analysis procedure, to the filtering of only 7 basis vectors. Another advantage is that a *small codebook size* satisfies for the storage of only 7 basis vectors. The basis vectors contain samples from a stochastic process with a zero-mean Gaussian distribution. The VSELP uses two of these codebooks which are searched one after the other, in the sequential approach.

The fourth innovation applied is the *adaptive post-filter* which enhances the synthetic output speech quality by attenuating coding noise in the inter-formant spectral regions. The spectral regions between adjacent harmonics of voiced speech are also attenuated, but not in the output speech. Instead, a *periodicity enhancement* is performed on the fixed-codebook excitation signal so that discontinuities introduced into the waveform are largely removed by the short-term synthesis filter.

The PDC (Personal Digital Cellular) full-rate standard, set in 1990 by the RCR (Research & development Center for Radio systems) of Japan, incorporates a speech coder which is a slightly modified version of the above VSELP coder. Its lower bit rate of 6.7 kbit/s is mainly obtained by using a single fixed codebook instead of two.

In 1991, the DoD defined a standard CELP at 4.8 kbit/s for military secure communication purposes [8]. For the first time in a standard CELP coder, *subsample resolution of the lag L* is used, which improves the quality of high-pitch voiced speech such as in some female and children's speech. The fixed codebook contains sparse, overlapping, *ternary-valued*, pseudo-randomly generated codewords. These properties firmly support complexity reduction. The ten short-term prediction coefficients are transmitted as uniformly *scalar quantized LSPs*, using 34 bits per

20-ms segment.

Also in 1991, the ITU-T standardized a *low-delay* (LD) CELP for wired-network applications at a bit rate of 16 kbit/s [10]. This LD-CELP features 50th-order *backward-adaptive short-term linear prediction* on the basis of the synthetic speech signal instead of the speech signal itself. Because the synthetic speech signal is generated both in the encoder and the decoder, the prediction coefficients are not transmitted. In addition, the coder makes use of a *backward-adaptive gain* parameter for the excitation signal. As a result, this parameter does not need to be transmitted as well and the full transmission capacity of 16 kbit/s is used for the excitation signal. The coder has no explicit pitch prediction. The segment length is only five samples and the intrinsic coding delay consequently amounts to the corresponding time duration of such a segment. Other contemporary CELP coders compute their short-term prediction coefficients in a forward way from segments with a duration in the order of 20 ms, giving rise to an algorithmic coding delay of the same duration.

Only 2 years later, in 1993, a PDC half-rate coding standard was completed [12]. The coder, called PSI-CELP (Pitch Synchronous Innovation), is a CELP variant with a pitch synchronised excitation source. Sometimes, the excitation from the fixed codebook is also referred to as "innovation" since this represents the unpredictable part of the speech signal, the final prediction residual after STP and LTP.

Apart from the adaptive codebook, the PSI-CELP uses a fixed codebook specifically intended for non-periodic speech parts and for periodic speech parts there are two other, stochastic, codebooks. The selected excitation vectors from these stochastic codebooks are each reduced in length to one pitch period and then *pitch synchronously repeated*. This feature shows some resemblance with the periodicity enhancement of the VSELP. The latter two codebooks are designed according to a *conjugate structure* [13] which enhances robustness with respect to errors in the transmission code. The low bit rate of 3.45 kbit/s is made possible by the use of 10-ms subsegments, thereby reducing the update rate of most transmission parameters by a factor of two with respect to contemporary standard coders. In addition, the codebooks are very small. The fixed codebook has a 5-bit code and the two stochastic codebooks have a 4-bit code each.

The ten short-term prediction coefficients are quantized on the basis of *vector quantization* (VQ), using *matrix*-VQ. This consists of *split*-VQ,

in which a separate vector quantizer is used for each subset (in this case 2, 4, 4) of the 10 coefficients, and *multistage*-VQ in each split-branch, albeit that the multistage-VQ is used in only one of the split-branches. In multistage-VQ the difference between the original vector and the vector-quantized vector is again subject to vector quantization. The result is that for the representation of the tenth-order set of LSPs only 23 bits are used. Despite the low bit rate of the coder this set is updated every 20 ms, as usual. This is the first time that VQ of the short-term prediction parameters is applied in a standard coder, although VQ in speech coding was proposed in the early eighties, already. The state of the art in VQ of the prediction coefficients in 1993 is found in [14].

The TIA standard speech coder IS-96 for the CDMA mobile radio system IS-95, also from 1993, is known as QCELP where Q stands for the company it originates from [17]. It is the first time that *variable bit rate* is applied. Variable bit rate can easily be utilized in a CDMA environment without the use of complicated time/frequency slot allocation as in TDMA. The coder operates at a bit rate of 8 kbit/s (full rate or rate 1) for active speech, at 4 or 2 kbit/s (rate 1/2 and rate 1/4) for background noise, and at 1 kbit/s (rate 1/8) for silence. The average bit rate of about 4 kbit/s is significantly lower than the full rate, while the speech quality essentially remains at the level of an 8 kbit/s coder. Application of other notable innovations in this coder include a one-dimensional *circular* fixed codebook in the excitation function and *differentially coded* LSPs (on the frequency axis) using 40 bits at rate 1, 20 bits at rate 1/2 and 10 bits at the rates 1/4 and 1/8. A full-duplex implementation runs on a single DSP, saving power dissipation with respect to continuous full rate operation, because the lower-rate programs are less complex than the full-rate program.

The half-rate GSM coder, proposed in 1993 [15], is a modified version of the VSELP coder at a bit rate of 5.6 kbit/s. The main features of this coder [11] are harmonic noise weighting, subsample resolution of the pitch lag and multi-mode (four levels of voicing) classification and treatment of the speech signal.

In addition use is made of a *hybrid search strategy for the lag* of the adaptive codebook, on the basis of open-loop pre-selection and closed-loop fine-tuning. This reduces complexity. In addition, vector quantization (VQ) of the short-term prediction coefficients is utilized. The coder applies split-VQ using 3 independent codebooks to 10 reflection coefficients in total. This results in a bit rate of 28 bits per 20-ms segment.

In 1995, the ITU-T adopts an 8 kbit/s standard [30]. The coder is a conjugate-structure *algebraic* CELP (CS-ACELP). The new features in this CELP variant are the segmentation and *predictive VQ scheme* for the LPC part resulting in a bit rate of 18 bits per 10 ms for 10 LSPs and an algorithmic coding delay of only 15 ms, but above all, the algebraic structure of the innovation source.

The innovation source is not a codebook, but basically a pulse sequence which is generated on the fly and not stored, just as in MPE and RPE. The structure of the pulse sequence prescribes a ternary-valued sequence in which a few pulses are placed which have normalised amplitudes of +1 or -1. The remaining time positions all have zero amplitude. The time positions of the pulses are constrained to a scheme which could be classified as something in between of MPE and RPE. In fact, there are a small number of grids (5) corresponding to decimation 5 and to each of the first three grids one binary pulse (+1 or -1) is assigned to the optimum position on the grid, and a single pulse is allocated to one of the remaining two grids. The optimum positions are determined by minimum perceptually weighted coding error, again. The ACELP algorithm also provides a computationally efficient way to determine these positions. The ACELP coding paradigm has turned out to be very successful. As we will see, it has been applied in several later standard coders.

Another ITU-T standard speech coder, also from 1995, operates at 5.3 or 6.3 kbit/s [16]. This coder, also referred to as the multimedia coder because it was meant to be used in teleconferencing (see Section 1.4), has little innovative features. The low rate operates on the basis of state-of-the-art ACELP with four grids (decimation 4) and 1 binary pulse per grid. The high rate works with decimation 2 grids and 6 binary pulses per grid for even subsegments and 5 for odd subsegments. The relatively low bit rate is made possible by using long segments (30 ms) and subsegments (7.5 ms). The algorithmic delay is as high as 37.5 ms.

In an Internal Philips Research Report [19] from 1995, a variable bit rate CELP coder with an *RPE-based innovation* source using binary pulses is described. The intended application was in digital telephone answering machines. The coder distinguishes between voiced speech and non-voiced speech (unvoiced and pauses). In the voiced mode the bit rate of the coder is 6.45 kbit/s and in the non-voiced mode 1.45 kbit/s. In the target application the average bit rate amounts to approximately 4.8 kbit/s. Much attention has been paid to make the algorithm suitable

for implementation on a single low-power DSP. On a proprietary DSP, the coder took 11 MIPS and the decoder 2 MIPS. The used ROM space was 6K words and the RAM space 2K words.

In 1995, the TIA proposed an enhanced variable-bit-rate codec (EVRC) for CDMA applications, described in its interim standard IS-127 [24]. The EVRC is an RCELP (Relaxed-CELP) [18], at a bit rate of 8.55 kbit/s for rate 1, 4 kbit/s for rate 1/2 and 0.8 kbit/s for rate 1/8. Rate 1/4 is not used, anymore. The main feature of the RCELP is a *generalized analysis-by-synthesis scheme* that involves a preprocessing functionality which modifies the pitch contour of the incoming speech signal into a perceptual equivalent in such a way that the requirements upon the adaptive codebook of the following ACELP coder are significantly relaxed. Thus, the transmitted lag-parameter uses only 12 bits per 20-ms segment, while in other coders the subsample-resolution lag-parameter requires 8 bit per 5-ms subsegment, typically.

After 1988, the year of completion of the first standard in public mobile telephony, the year 1995 forms a second landmark in this overview. In this period the CELP method of coding had become mature and applied widely. Several surveys on the area became available, with the book Speech Coding and Synthesis of Kleijn and Paliwal (eds.) [23] as a highlight. All above mentioned techniques and innovations are included in it. Other informative overviews are found in [20] and [22].

In the period after 1995 more variants based on the CELP coding paradigm have been developed and applied. In the fall of 1995, the ETSI issued the ACELP-based enhanced full-rate (EFR) speech coder for its GSM system [26] giving enhanced speech quality at a similar bit rate as the full-rate coder. The enhanced quality had been enabled by the developments in the preceding years. In 1996, the same coder is also applied as the TIA standard IS-641 in the North American TDMA digital cellular system TIA IS-136 [25]. The full-rate coder in this system was already for some time the IS-54 VSELP.

In 1999 the *adaptive multi-rate* (AMR) GSM coder [31] is standardized by the ETSI as part of the GSM-AMR system. The feature of this system is that the constant bit rate of 22.4 kbit/s in a GSM voice channel is adaptively divided between the speech coder and the channel coder, dependent on the quality of the voice channel. The better the radio channel, the higher the bit rate for the speech coder. The control system realizing this feature forms the main innovation in the AMR system. Measuring and signalling of the radio-channel quality is necessarily

realized within the voice-band. It is based on a speech quality measure derived from the error protection/correction system (the Viterbi channel decoder). The speech coder is ACELP/EFR based and the variable bit rate is mainly realized by varying the degree of freedom in the algebraic excitation generator. The bit rate varies between 4.75 and 12.2 kbit/s. Our own RPE-based candidate, which did not survive in the standardization contest for AMR, has been described in [50] and [51]. The paradigm of RPE grids, however, will live on in the ACELP grids.

Also in 1999, the MPEG-4 speech coding standards are approved and issued [33]. These standards are primarily developed for internet use and the focus has been on flexibility, mainly in terms of *scalability* of the bit stream. They have also been mentioned in Section 1.4. Our own contribution to this standard was presented in [48] and a review of this work in a wider scope was presented in [49].

A compact overview of the state of the art in 2000, including the above mentioned coders, is found in the book of Goldberg and Riek [34]. By this time another period in the development of the CELP coding has ended, characterised by the fact that the basic standards have been set and that future technological developments are in a lower gear. The application of speech coding technology, however, keeps on growing.

In 1998, five standards developing organizations in Europe, the United States, Japan, Korea and China agreed to create the Third Generation Partnership Project (3GPP) on mobile communications. In 2000, the ETSI Special Mobile Group was dissolved and the specification of GSM evolution, including ETSI's universal mobile telecommunications system UMTS, was transferred to 3GPP [35]. In 1999, four standards developing organizations from the USA, Japan, Korea and China created 3GGP2 to specify and maintain globally applicable non-GSM systems such as the CDMA-based radio systems IS-95 and cdma2000 [36]. In 2000, the TIA and 3GPP2 issued a new speech coding standard for CDMA applications as a successor of the EVRC coder, called the *selectable mode* vocoder (SMV) [37], [38]. This coder uses variable rate, just as the EVRC, depending on the time varying properties of the input speech, including the quarter rate again. On top of that, three modes with different average bit rates can be selected independent of the speech signal, a similar functionality as in GSM-AMR. The kernel of the speech coder is based on the so called "extended" CELP (eX-CELP) coding method [39]. The main feature of the eX-CELP coding method is a *hybrid closed-loop/open-loop* approach combined with a

careful selective use of them guided by the varying properties of input speech signals. The heart of the eX-CELP is based on a generalized analysis-by-synthesis coding paradigm like RCELP. The eX-CELP has also been proposed for the ITU-T 4 kbit/s coding standard [41], which is pending since 1994 [40]. Although this coder could meet the majority of the requirements, there is still no candidate meeting all requirements. Another candidate which passes seven out of eight requirements is a hybrid MELP/ CELP coder [42], [43]. The MELP coding paradigm is addressed in Section 5.3.

In 2001, 3GPP proposed an *AMR wide-band* coder (AMR-WB) [46]. The coder is based on EFR and modified for 16 kHz sampling rate. The *variable-rate multi-mode wide-band* (VMR-WB) speech coding standard for cdma2000 [47] application was issued in 2003 by 3GPP2. The VMR-WB was built on QCELP and EVRC paradigms and, in addition, it has a mode that is inter-operable with AMR-WB. This worldwide *inter-operable wide-band standard* is now being considered to become an ITU-T standard (G.722.2). Our own activities in the field of wide-band coding were focused on *parametric coding of the upper-band* (4–8 kHz) [52] and the generation of an artificial upper-band from only the narrow-band speech signal [53], also referred to as *bandwidth extension.*

5.3 Developments in vocoders

The state of the art in vocoders in 1994 is reflected in the elaborate survey of Jaskie and Fette [21]. Tens of existing vocoders pass in review, most of them using LPC in spectral modelling.

A milestone in the development of LPC-vocoders is the mixed-excitation LP (MELP) vocoder at 2400 bits/s [27], [28], which was chosen in 1996 by the DoD as the new Federal Standard replacing the old LPC-10 standard FS1015 from 1982. It incorporates the usual LPC synthesis filter and characteristic for this vocoder are the *mixed excitation* using pulses and noise, the flexibility in using *periodic or aperiodic pulses*, the *adaptive spectral enhancement* which enhances the formants and in addition corrects for possible spectral tilt, the use of *pulse dispersion filtering* to reduce the well-known "buzziness" and the application of *Fourier magnitude modelling* increasing the spectral accuracy. It was reported that it has been implemented on a single-chip floating-point DSP using 14 MIPS for the encoder and 6 MIPS for the decoder. Memory usage was also reported. The implementation used 9

kwords for the program, 6 kwords for the tables and 6 kwords of RAM.

In addition to the LPC vocoders, the class of sinusoidal vocoders has been very successful. The basic idea of sinusoidal coding was proposed by Dudley already in 1940, as described in Section 2.3.3. The first standard based on this coding paradigm was the IMBE vocoder [9] for the Inmarsat application, as already mentioned in Chapter 1. This standard was set in 1990. An own sinusoidal vocoder, called "Harmony-1", has been described in an Internal Philips Research Report [29] from 1997. It was especially designed for telephone answering machines. It uses *LPC to encode the amplitudes* of the speech harmonics and a *joint optimisation* of these amplitudes and the pitch. Despite the use of an LPC synthesis filter this kind of vocoder *does not suffer from buzziness* thanks to the phase-continuous synthesis of the individual harmonics. Phase continuity is the continuation of each individual harmonic without any phase jumps in the course of time. The report describes a variable bit rate coding scheme that is capable of reconstructing natural sounding speech at an average bit rate of 3.2 kbit/s. A special feature of this sinusoidal coder is that it allows to play-back recorded messages at a lower or higher speed with the same pitch contour as the original message. It has been implemented on a DSP where the encoder takes about 12 MIPS and the decoder 6 MIPS. It uses 5 kwords of ROM and about 1.1 kwords of RAM. Several millions of these coders have been applied in telephone answering machines and toys, all over the world. It was probably the first time that a vocoder was so widely spread over the globe. Other own work related to his field included pitch synchronous interpolation [54] and the time warping of speech signals [56].

Another sinusoidal vocoder which has been standardized in MPEG-4 in 1999 is described in [32]. The vocoder is called HVXC (from harmonic vector excitation coding) and it has a bit rate of 2 kbit/s. It mainly differs from Harmony-1 by the use of spectral-amplitude parameters of the prediction residual, in addition to the LPC parameters like in the MELP vocoder, and they are conveyed to the decoder in vector-quantized form.

5.4 Non-speech-specific coding

Sinusoidal coding is also used for the coding of non-speech signals. These signals mostly concern audio signals. The main difference with speech coding is that the sinusoids representing spectral components in the

signal do not belong to a single harmonic complex. This necessitates specific coding techniques to maintain efficient low-bit-rate coding. All frequencies have to be coded individually and transmitted. One instrument in efficient frequency coding is linking. This implies that for subsequent signal segments sinusoids from one segment are linked to sinusoids of the other segment, thus creating tracks in the course of time. These tracks of time-varying frequencies should represent original tones thus enabling phase continuity in the reconstruction of these tones and differential coding of the relatively slowly varying frequencies. An MPEG-4 coder based on this principle is addressed in the next section.

In an experiment to use this technique for telephony, interesting results have been obtained. A narrow-band sinusoidal codec has been designed [57], [61], at a bit rate of 12.2 kbit/s, the same bit rate as that of the GSM-EFR coder. The objective was to improve the sound quality for non-speech signals (music) and to maintain the quality for speech signals, relative to the performance of the EFR coder. The results showed that the sinusoidal coder performed significantly better for most music signals than the EFR coder. For speech, however, the performance was slightly worse than the EFR coder.

Some other own contributions in this area concern an analysis-by-synthesis linking technique [55], the efficient transmission of phase information [58] enabling signal reconstruction with original phases which also reconstructs the original waveform, and the use of RPE-based residual coding to enhance the sound quality of sinusoidal coders [59], [60].

Another non-speech-specific coder concerns the wide-band transform coder of the ITU-T standard G.722.1 [45], [44] from 1999. It has a bit rate of 24 kbit/s or 32 kbit/s. This coder shows more conformity with audio coding techniques relying only on perceptual models, than with speech coding technology.

5.5 Outlook

The trends in speech coding are depicted in Figure 5.1. The figure shows the locations of a selection of milestone coders, most of which are standards, in a logarithmic-bit-rate versus linear-time plot. The location on the time axis is the approximate year of introduction of each coder. The main features of these coders have been addressed before, apart from the wide-band speech coder ACELP/TCX (Transform-Coded

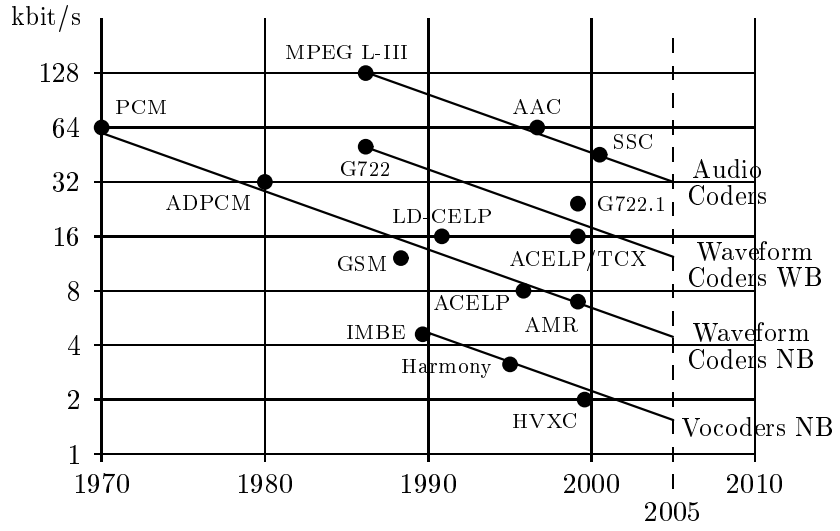

Figure 5.1: Trends in speech and audio coding.

eXcitation) [62], which was a candidate in the contest for the ITU-T G722.1 standard, and the audio coders MPEG L-III (better known as MP3) and AAC (Advanced Audio Coder). These audio coders will not be specified here, but the reader is referred to [63]. The indicated SSC (SinuSoidal Coder) is an early version of the recent MPEG-4 amendment (see below) on parametric audio coding. For each of the speech coding fields of narrow-band vocoders, narrow-band waveform coders and wide-band waveform coders, a trend line is rendered. A trend line for audio coding is also shown. The result of plotting these trend lines in such a way is striking. They are practically parallel and they seem to be almost equidistant. The picture was composed for the first time in early 2001 and used for predicting the situation in 2005–2010. Now, in 2005, we have to conclude that this goal has been served remarkably well. In narrow-band waveform coders, the ITU-T 4 kbit/s standard has become quite close albeit that it is not yet completed. In wide-band waveform coders the AMR-WB and VMR-WB standards, as mentioned in the previous section, have been realized at the indicated bit rates. In narrow-band vocoders no further activity has been observed after 2000, but the indicated bit rate of about 1600 bit/s in 2005 is certainly feasible.

The audio line, finally, gives a quite accurate indication of the present state of the art in audio coding as reflected by the MPEG-4 so-called "Amendments" High-Efficiency AAC (HE-AAC) [64], which has also been adopted by 3GPP under the name "aacPlus", and the parametric audio coder based on the sound objects of transients, sinusoids and noise [65].

Based on technological insights, evaluation of the world-wide academic and industrial efforts and sensing the activities at international conferences and workshops, it is concluded that it is likely that the trend lines of speech coding as well as the trend line of audio coding, will *not* be prolonged from, say, 2005 onwards. Instead, they will probably *saturate*. This event would mark the end of an era and terminate the fruitful harvesting period in speech and audio coding concentrated around the 1990s.

References

1. T. Aoyama, W.R. Daumer, G. Modena (Guest Editors), Issue on voice coding for communications, *IEEE Journal on Selected Areas in Communications*, Volume 6, No.2, February 1988.
2. I.M. Trancoso, B.S. Atal, Efficient procedures for finding the optimum innovation in stochastic coders, *IEEE International Conference on Acoustics Speech and Signal Processing*, Tokyo, April 1986, pp. 2375–2378.
3. P. Kroon, E. F. Deprettere, A class of analysis-by-synthesis predictive coders for high quality speech coding at rates between 4.8 and 16 kbits/s, *IEEE Journal on Selected Areas in Communications*, Volume 6, No.2, February 1988, pp. 353–363.
4. P. Mermelstein, G.722, a new CCITT coding standard for the digital transmission of wideband audio signals, *IEEE Communications Magazine*, Volume 26, No.1, Januari 1988, pp. 8–15.
5. X. Maitre, 7 kHz audio coding within 64 kbit/s, *IEEE Journal on Selected Areas in Communications*, Volume 6, No.2, February 1988, pp. 283–298.
6. I. A. Gerson and M. A. Jasiuk, Vector sum excited linear prediction (VSELP) speech coding at 8 kbps, *IEEE International Conference on Acoustics Speech and Signal Processing*, Albuquerque, April 1990, pp. 461–464.
7. M. H. Sunwoo and S. Park, Real-time implementation of the VSELP on a 16-bit DSP chip, *IEEE Transactions on Consumer Electronics*, Volume 37, No.4, November 1991, pp. 772–782.
8. J.P. Campbell Jr., T.E. Tremain, V.C. Welch, The DoD 4.8 kbps standard (proposed federal standard 1016), *Advances in speech coding, B.S. Atal, V. Cuperman, a. Gersho (eds.)*, Kluwer, Boston, 1991.
9. J.C. Hardwick, J.S. Lim, The application of the IMBE speech coder to mobile communications, *IEEE International Conference on Acoustics Speech and Signal Processing*, Toronto, Canada, May 1991, pp. 249–252.
10. J-H. Chen, R. Cox, Y-C. Lin, N. Jayant, M.J. Melchner, A low-delay CELP coder for the CCITT 16 kbit/s speech coding standard, *IEEE Journal on Selected Areas in Communications*, Volume 10, No.5, June 1992, pp. 830–849.

11. I.A. Gerson and M. A. Jasiuk, Techniques for improving the performance of CELP-type speech coders, *IEEE Journal on Selected Areas in Communications*, Volume 10, No.5, June 1992, pp. 858–865.
12. T. Ohya, H. Suda, T. Miki, 5.6 kbit/s PSI-CELP of the half-rate PDC speech coding standard, *IEEE 44th Vehicular Technology Conference*, 1994, pp. 1680–1684.
13. T. Moriya, Two-channel conjugate vector quantizer for noisy channel speech coding, *IEEE Journal on Selected Areas in Communications*, Volume 10, No.5, June 1992, pp. 866–874.
14. K.K. Paliwal, B.S. Atal, Efficient vector quantization of LPC parameters at 24 bits/frame, *IEEE Transactions on Speech and Audio Processing*, Volume 1, No.1, Januari 1993, pp. 3–14.
15. I. A. Gerson and M. A. Jasiuk, A 5600 bps VSELP speech coder candidate for half-rate GSM, *IEEE Workshop on Speech Coding*, Sainte-Adèle, Canada, October 1993, pp. 43–44.
16. Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, *ITU-T Recommendation* G.723.1.
17. A. DeJaco, W. Gardner, P. Jacobs, C. Lee, QCELP: The North American CDMA digital cellular variable rate speech coding standard, *IEEE Workshop on Speech Coding*, Sainte-Adèle, Canada, October 1993, pp. 5–6.
18. W.B. Kleijn, P. Kroon, D. Nahumi, The RCELP speech-coding algorithm, *European Transactions on Telecommunications*, Volume 5, No.5, September-October, 1994, pp. 39/573–48/582.
19. R.J. Sluijter, E. Kathmann, A versatile speech coder for storage applications, *Philips Research Internal Report*, TN 287/95, 1995.
20. A. Gersho, Advances in speech and audio coding, *Proceedings of the IEEE*, Volume 82, No.6, June 1994, pp. 900–918.
21. C. Jaskie, B. Fette, A survey of low bit rate vocoders, *DSP & Multimedia Technology*, April 1994, pp. 26–40.
22. R.J. Sluijter, F. Wuppermann, R. Taori and E. Kathmann, State of the art and trends in speech coding, *Philips Journal of Research*, Volume 49, No.4, 1995, pp. 455–488.
23. W.B. Kleijn and K.K. Paliwal (eds.), *Speech Coding and Synthesis*, Elsevier, Amsterdam, 1995.
24. Enhanced variable rate codec, speech service option 3 for wideband spread spectrum digital systems, *TIA/EIA/IS-127 Standard*, September 9, 1996.

25. T. Honkanen, J. Vainio, K. Järvinen, P. Haavisto, Enhanced full rate speech codec for the IS-136 digital cellular system, *IEEE International Conference on Acoustics Speech and Signal Processing*, Munich, Germany, April 1997, pp. 731–734.
26. K. Järvinen, J. Vainio, K. Kapanen, T. Honkanen, P. Haavisto, R. Salami, C. Laflamme, J-P. Adoul, GSM enhanced full rate speech codec, *IEEE International Conference on Acoustics Speech and Signal Processing*, Munich, Germany, April 1997, pp. 771–774.
27. A. McCree, K. Truong, E.B. George, T.P. Barnwell, V. Viswanathan, A 2.4 kbit/s MELP coder candidate for the new U.S. federal standard, *IEEE International Conference on Acoustics Speech and Signal Processing*, Atlanta, Georgia, May 1996, pp. 200–203.
28. L.M. Supplee, R.P. Cohn, J.S. Collura, MELP: the new federal standard at 2400 bps, *IEEE International Conference on Acoustics Speech and Signal Processing*, Munich, Germany, April 1997, pp. 1591–1594.
29. R. Taori, R.J. Sluijter, A.J. Gerrits, Harmony-1: a versatile low bit rate speech coding system, *Philips Research Internal Report*, TN 157/97, 1997.
30. R. Salami, C. Laflamme, J-P. Adoul, A. Kataoka, S. Hayashi, T. Moriya, C. Lamblin, D. Massaloux, S. Proust, P. Kroon, Y. Shoham, Design and description of CS-ACELP: a toll quality 8 kb/s speech coder, *IEEE Transactions on Speech and Audio Processing*, Volume 6, No.2, March 1998, pp. 116–130.
31. E. Ekudden, R. Hagen, I. Johansson, J. Svedberg, The adaptive multi-rate speech coder, *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999, pp. 117–119.
32. M. Nishiguchi, A. Inoue, Y. Maeda, J. Matsumoto, Parametric Speech Coding - HVXC at 2.0-4.0 kbps, *Proceedings of the IEEE Workshop on Speech Coding*, Porvoo, Finland, 1999, pp. 84–86.
33. K. Brandenburg, O. Kunz and A. Sugiyama, MPEG-4 Natural Audio Coding, *Signal Processing: Image Communication*, Tutorial Issue on MPEG-4 Standard, Volume 15, Nos.4-5, January 2000, pp. 423–444.
34. R. Goldberg, Lance Riek, *A practical handbook of speechcoders*, CRC Press, Boca Raton, Florida, 2000.
35. J.F. Huber, D. Weiler, H. Brand, UMTS, the mobile multimedia

vision for IMT-2000: a focus on standardization, *IEEE Communication Magazine*, September 2000, pp. 129–136.

36. M. Zeng, A. Annamalai, V.K. Bhargava, Harmonization of global third-generation mobile systems, *IEEE Communication Magazine*, December 2000, pp. 94–104.
37. Y. Gao, E. Shlomot, A. Benyassine. J. Thyssen, H-y. Su, C. Murgia, The SMV algorithm selected by the TIA and 3GPP2 for CDMA applications, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, May 2001.
38. S.C. Greer, A. DeJaco, Standardization of the selectable mode vocoder, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, May 2001.
39. Y. Gao, A. Benyassine. J. Thyssen, H-y. Su, E. Shlomot, eX-CELP: a speech coding paradigm, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, May 2001.
40. S. Dimolitsas, C. Ravishankar and G. Schroeder, Current Objectives in 4-kb/s Wireline-Quality Speech Coding Standardization, *IEEE Signal Processing Letters*, November 1994, pp. 157–159.
41. J. Thyssen, Y. Gao, A. Benyassine, E. Shlomot, C. Murgia, H-y. Su, K. Mano, Y. Hiwasaki, H. Ehara, K. Yasunaga, C. Lamblin, B. Kovesi, J. Stegmann, H-G. Kang, A candidate for the ITU-T 4 kbit/s speech coding standard, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, Utah, May 2001.
42. A. McCree, J. Stachurski, T. Unno, E. Ertan, E. Paksoy, V. Viswanathan, A. Heikkinen, A. Rämö, S. Himanen, P. Blöcher, O. Dressler, A 4 kb/s hybrid MELP/CELP speech coding candidate for ITU standardization, *IEEE International Conference on Acoustics Speech and Signal Processing*, Orlando, Florida, May 2002, pp. I629–I632.
43. J. Stachurski, A. McCree, V. Viswanathan, A. Heikkinen, A. Rämö, S. Himanen, P. Blöcher, Hybrid MELP/CELP coding at bit rates from 6.4 to 2.4 kb/s, *IEEE International Conference on Acoustics Speech and Signal Processing*, Hong Kong, 2003, pp. II153–II156.
44. A. Crossman, A fixed bit rate wideband codec for use in ISDN videotelephony, *IEEE Workshop on Speech Coding*, Anapolis, Maryland, September 1995, pp. 29–30.

45. Coding at 24 and 32 kbit/s for hands-free operation in systems with low frame loss, *ITU-T Standard* G.722.1, September 1999.
46. AMR wideband speech codec, *3GPP Standard* TS 26.190 (V1.0), March 2001.
47. M. Jelinek, R. Salami, S. Ahmadi, B. Besette, P. Gournay, C. Laflamme, On the architecture of the CDMA2000 variable-rate multimode wideband (VMR-WB) speech coding standard, *IEEE International Conference on Acoustics Speech and Signal Processing*, Montreal, Canada, May 2004, pp. I281–I284.
48. R. Taori, R.J. Sluijter and A.J. Gerrits, On scalability in CELP coding systems, *IEEE Workshop on Speech Coding for Telecommunications*, Pocono Manor, Pennsylvania (USA), 1997, pp. 67–68.
49. R.J. Sluijter, Scalable speech coding for multimedia, (invited paper), *Proc. of IEEE Benelux Signal Processing Chapter and Flandres Language Valley Education "DSP aspects of speech processing"*, Ieper, Belgium, December 1999.
50. A. Gerrits, R. Taori, R.J. Sluijter, R.C.E. Heijmans, Speech coding for GSM-AMR, *Philips Research Internal Report*, TN 062/99, February 1999.
51. A. Gerrits, A. Koppelaar, R. Taori, R. Sluijter, C. Baggen, E. Hekstra-Nowacka, Proposal for an adaptive multi-rate coder for GSM, *Proceedings of the 20th Symposium on Information Theory in the Benelux*, Haasrode, Belgium, 1999.
52. R. Taori, R.J. Sluijter, A.J. Gerrits, Hi-BIN: an alternative approach to wideband speech coding, *IEEE International Conference on Acoustics Speech and Signal Processing*, Istanbul, Turkey, 2000.
53. S. Chennoukh, A. Gerrits, G. Miet, and R. Sluijter, Speech enhancement via frequency bandwidth extension using line spectral frequencies, *IEEE International Conference on Acoustics Speech and Signal Processing*, Salt Lake City, USA, 2001.
54. R. Taori, R.J. Sluijter and E. Kathmann, Speech compression using pitch synchronous interpolation, *IEEE International Conference on Acoustics Speech and Signal Processing*, 1995, Detroit.
55. R. Taori and R.J. Sluijter, Closed-loop tracking of sinusoids for speech and audio coding, *Proceedings of the 1999 IEEE Workshop on Speech Coding*, Porvoo, Finland, 1999, pp. 1–3.
56. R.J. Sluijter and A.J.E.M. Janssen, A time warper for speech signals, *Proceedings of the 1999 IEEE Workshop on Speech Coding*, Porvoo, Finland, 1999, pp. 150–152

57. G. Hotho and R. Sluijter, A low bit rate audio and speech coder for narrowband signals, *1st Benelux Workshop on Model based Processing and Coding of Audio*, Leuven, Belgium, November 15, 2002.
58. A.C. den Brinker, A.J. Gerrits, R.J. Sluijter, Phase transmission in a sinusoidal audio and speech coder, *Audio Engineering Society, AES 115th Convention*, Paper 5983, 10-13 October, 2003, New York USA.
59. F. Riera-Palou, A.C. den Brinker, A.J. Gerrits and R.J. Sluijter, Improved optimisation of excitation sequences in speech and audio coders, *Electronics Letters*, Volume 40 (8), 15th April 2004, pp. 515–517.
60. F. Riera-Palou, A.C. den Brinker, A.J. Gerrits and R.J.Sluijter, Improved optimisation of excitation sequences in speech and audio coders, *Proc. 4th IEEE Benelux Signal Processing Symposium*, Hilvarenbeek (NL), April 15-16, 2004, pp. 53–56.
61. G. Hotho and R. Sluijter, A narrowband low bit rate sinusoidal audio and speech coder, *Proceedings EUSIPCO 2004, XII, European Signal Processing Conference*, 6-10 September 2004, Vienna, Austria, pp. 1661–1664.
62. B. Besette, R. Salami, C. Laflamme, R. Lefebvre, A wideband speech and audio codec at 16/24/32 kbit/s using hybrid ACELP/TCX techniques, *IEEE Workshop on Speech Coding*, Porvoo, Finland, June 1999, pp. 7–9.
63. T. Painter and A. Spanias, Perceptual coding of digital audio, *Proceedings of the IEEE*, Volume 88, No.4, April 2000, pp. 451–512.
64. ISO/IEC, Information technology - coding of audio-visual objects. Part3: Audio, Amendment1: High-efficiency AAC, *ISO/IEC IS14496-3:2001/AMD1:2004*,July 2004.
65. ISO/IEC, Information technology - coding of audio-visual objects. Part3: Audio, Amendment2: Parametric coding of high quality audio, *ISO/IEC IS14496-3:2001/AMD2:2004*,July 2004.

Appendix A

The plane wave equations

Consider the cross-sectional slice of a gas in a lossless uniform tube, with a fixed mass m being part of a plane wave, as shown in Figure A.1. The area of the slice is $\mathcal{A}$. The pressure on the left wall of the slice is $P(x,t)$

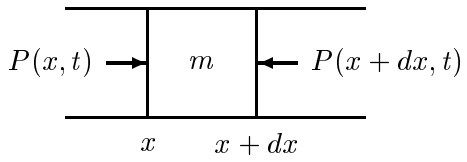

Figure A.1: Slice of gas with a fixed mass m in a plane wave.

and on the right wall $P(x+dx,t)$. So, the net force F on the slice is given by

$$F = \mathcal{A}P(x,t) - \mathcal{A}P(x+dx,t) = -\mathcal{A}\frac{\delta p(x,t)}{\delta x}dx \ , \qquad \text{(A.1)}$$

in which $P(x,t) = P_0 + p(x,t)$, and P_0 is the pressure of the gas in equilibrium. According to Newton's second law, the law of motion, the relation between F, m and the velocity of the slice $v(x,t) = u(x,t)/\mathcal{A}$, where u stands for volume velocity, is given by

$$-\mathcal{A}\frac{\delta p(x,t)}{\delta x}dx = \frac{m}{\mathcal{A}}\frac{\delta u(x,t)}{\delta t} \ . \qquad \text{(A.2)}$$

Substitution of $m = \varrho\mathcal{A}dx$, with ϱ the density of the gas, yields the first wave equation describing the inertia of the mass:

$$-\frac{\delta p(x,t)}{\delta x} = \frac{\varrho}{\mathcal{A}}\frac{\delta u(x,t)}{\delta t} \ . \qquad \text{(A.3)}$$

The density $\varrho = \varrho(x,t)$ is a function of x and t. This second order effect can often be neglected in this equation, however. In air, for example, its variation with respect to the atmospheric density ϱ_0 is very low. Only at very loud sound levels the variation can reach $10^{-3}\varrho_0$.

The second wave equation arises from the compressibility of the gas. If we consider a cross-sectional compartment of the tube with a fixed volume V_0 according to Figure A.2, we see that the inflow of gas at location x into this compartment during a small time dt amounts to $u(x,t)dt$. The outflow at location $x+dx$ during the same time dt amounts

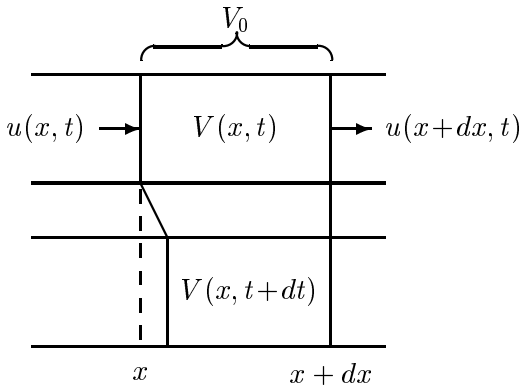

Figure A.2: Fixed volume of a gas V_0 in a plane wave at time t (top) and $t+dt$ (bottom).

to $u(x+dx,t)dt$. As a result, the total net inflow into the compartment with volume V_0 amounts to the volume:

$$V(x,t) - V(x,t+dt) = u(x,t)dt - u(x+dx,t)dt \;, \tag{A.4}$$

which can also be written as

$$-\frac{\delta V(x,t)}{\delta t}dt = -\frac{\delta u(x,t)}{\delta x}dxdt \;. \tag{A.5}$$

So, after the inflow, the compartment contains the compressed original volume of the gas plus the new inflow. Consequently, the pressure in the compartment has been increased. In order to relate this change in pressure to the change in volume of the original mass of gas, the gas law is used, and more specifically, the gas law for an adiabatic process (this

is the case in which there is no heat exchange between any partial compartment of the gas and its environment, which is a good representation of the situation because gasses are bad heat conductors):

$$PV^{\eta} = \text{constant} \ , \tag{A.6}$$

in which η is the adiabatic constant which depends on the kind of gas, $P = P_0 + p(x,t)$ and $V = V(x,t)$. Partial differentiation with respect to t of this expression yields

$$\frac{\delta V}{\delta t} = -\frac{V}{\eta P}\frac{\delta p(x,t)}{\delta t} \ . \tag{A.7}$$

If we combine this expression with Equation A.5, we get the second wave equation

$$-\frac{\delta u(x,t)}{\delta x} = \frac{\mathcal{A}}{\eta P}\frac{\delta p(x,t)}{\delta t} \ , \tag{A.8}$$

in which $V(x,t) = \mathcal{A}dx$ has been substituted which equality was already implicitly assumed as appears from Figure A.2 (top). In this second wave equation it must be noted that P is not constant but a function of x and t, just as was the case for ϱ in the first wave equation. Also here, P can often very well be approximated by P_0 because only in very loud sounds $p(x,t)$ reaches $10^{-3}P_0$.

Remark on the propagation speed of sound

According to Equations 2.6 and 2.8 the propagation speed of sound in a gas was given by

$$c = \sqrt{\eta\frac{P}{\varrho}} \ . \tag{A.9}$$

Here, P and ϱ are both functions of x and t as we have seen above, but the quotient P/ϱ is constant, however. In order to clarify this we apply the universal gas law (which also holds for the adiabatic case) to the constant mass of gas in the volume $V(x,t)$:

$$PV = nRT \ , \tag{A.10}$$

in which R is the universal gas constant and T the absolute temperature. The volume V concerns a volume that contains n moles of the gas. One mole is M kg of a matter which has molecular weight M. Therefore the density of the gas is:

$$\varrho = \frac{nM}{V} \ . \tag{A.11}$$

By combining Equations A.10 and A.11 we find

$$\frac{P}{\varrho} = \frac{RT}{M} \,. \tag{A.12}$$

So, we infer that P/ϱ depends exclusively on the temperature because R and M are constants. In this way we can conclude

$$c = \sqrt{\eta \frac{P}{\varrho}} = \sqrt{\eta \frac{RT}{M}} \,. \tag{A.13}$$

For air $\eta = 1.4$. With $P = 10^5$ [Newton/m^2] and $\varrho = 1.28$ [kg/m^3] (both at 0^0C) we get $c_0 = 331$ [m/s]. With $R = 8.31451$ [Newton mK^{-1}mol^{-1}] and $M = 28.964\ 10^{-3}$ [kg mol^{-1}] and $T = 273$ [K] we also get $c_0 = 331$ [m/s]. For body temperature (37^0C) c_{37} becomes 354 [m/s].
Oxygen: $\eta = 1.4$, $\varrho = 1.43$ (at 0^0C): $c_0 = 313$ [m/s].
Helium: $\eta = 1.66$, $M = 4.003\ 10^{-3}$: $c_0 = 970$ [m/s].

In Appendix B of Fletcher's book on speech and hearing from 1929 [1], similar wave equations and expressions for the propagation speed of sound can be found. More recent references giving similar expressions are given in [2] and [3].

References

1. H. Fletcher, *Speech and Hearing*, D. van Nostrand Company, Inc., New York, 1929, Appendix B.
2. L.L. Beranek, *Acoustics*, McGraw-Hill Book Company, Inc., New York, 1954.
3. E. Skudrzyk, *The foundations of Acoustics*, Springer-Verlag, Wien, 1971.

Appendix B

The duality of acoustic tubes

In the discrete-time domain the transfer function of a network can be obtained by evaluating the z-transform $H_d(z)$ of the impulse response $h[n]$ at the unit circle $z = e^{j\omega\tau}$. By doing so, the z-transform turns into the discrete-time Fourier transform:

$$H_d(e^{j\omega\tau}) = \sum_{n=-\infty}^{\infty} h[n]e^{-j\omega n\tau} \ . \tag{B.1}$$

In the continuous-time domain the transfer function $H_c(\omega)$ can in the same way be found by taking the continuous-time Fourier transform of the impulse response, which obviously (from Figure 2.23) consists of impulses spaced by τ:

$$H_c(\omega) = \int_{-\infty}^{\infty} \sum_{n=-\infty}^{\infty} g(t)\delta(t - n\tau)e^{-j\omega t}dt \ , \tag{B.2}$$

where $\delta(t)$ stands for the Dirac impulse and $g(n\tau) = h[n]$. By interchanging the integration and summation operators and using the fact that for any finite signal $x(t)$

$$\int_{-\infty}^{\infty} x(t)\delta(t - t_0)dt = x(t_0) \ , \tag{B.3}$$

where t_0 stands for an arbitrary instant, we find

$$H_c(\omega) = \sum_{n=-\infty}^{\infty} h[n]e^{-j\omega n\tau} \ . \tag{B.4}$$

This shows clearly the duality of both versions of H.

Appendix C

Time-flip-and-shift in lattice filters

We will prove the relation

$$B_{m-1}(z) = z^{-m} A_{m-1}(z^{-1}) \tag{C.1}$$

according to Equation 2.53, which prescribes that $B_{m-1}(z)$ is obtained by a time flip of $A_{m-1}(z)$ and subsequently time shift of the result, by induction (after [1]). To this end we observe from Equation 2.53 and the signal-flow graph of $A(z)$ of Figure 2.28, that the equation is true for $m = 1$ and $m = 2$. For $m = 1$ we find indeed

$$A_0(z) = 1 \tag{C.2}$$

and

$$B_0(z) = z^{-1} = z^{-1} A_0(z^{-1}) \; . \tag{C.3}$$

For $m = 2$ we indeed find

$$A_1(z) = 1 + r_1 z^{-1} \tag{C.4}$$

and

$$B_1(z) = z^{-1}\{r_1 + z^{-1}\} = z^{-2} A_1(z^{-1}) \tag{C.5}$$

If the relation of Equation C.1 is true, then the relations

$$A_{m-1}(z^{-1}) = z^{m} B_{m-1}(z) \tag{C.6}$$

and

$$A_{m-1}(z) = z^{-m} B_{m-1}(z^{-1}) \tag{C.7}$$

are also true, and substitution of Equations C.1 and C.7 into Equation 2.52 and combination with Equation 2.51 finally yields

$$B_m(z) = z^{-(m+1)} A_m(z^{-1}) \; . \tag{C.8}$$

This proves the assumption of Equation C.1.

References

1. L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978, pp. 92–98, (Section 3.3.4: Transfer function of the lossless tube model).

Appendix D

Bandwidth expansion in LPC

In the applications of LPC a window is often used on the obtained a-parameters to avoid sharp resonances in the synthesis filter $1/A(z)$. This window is usually defined for the inverse filter $A(z)$. The transfer function of the inverse filter itself is given by

$$A(z) = 1 + \sum_{i=1}^{M} a_i z^{-i}, \tag{D.1}$$

where M stands for the order of the LPC. With the window this becomes

$$A_\gamma(z) = 1 + \sum_{i=1}^{M} (a_i \gamma^i) z^{-i}. \tag{D.2}$$

The time window applied to the impulse response

$$\alpha[n] = \begin{cases} 1 & , n = 0, \\ a_n & , n = 1, \ldots, M \end{cases} \tag{D.3}$$

of $A(z)$ is here given by γ^n , with $n = 0, 1, \ldots, M$, and it is a decaying window with $0 \leq \gamma \leq 1$. Application of the window function to direct-form realizations of $A(z)$ or $1/A(z)$ can be implemented by replacing the coefficients a_i by $a_i \gamma^i$. We can also rewrite Equation D.2 as

$$A_\gamma(z) = 1 + \sum_{i=1}^{M} a_i (z/\gamma)^{-i} = A(z/\gamma). \tag{D.4}$$

From this expression it can be seen that applying the window function to a lattice-form realization of $A(z)$ or $1/A(z)$ can be implemented by replacing every delay operator z^{-1} in the signal-flow graph by $\gamma.z^{-1}$.

Because multiplication of the impulse response by the window gives rise to a convolution of the corresponding frequency transforms in the frequency domain, it can be concluded that there is a bandwidth expansion of possible sharp valleys in $|A(z)|$ and consequently also of sharp resonances in $1/|A(z)|$. It can also be argued that if the zeroes of $A(z)$ are located at $z = p_i$ then the zeroes of $A(z/\gamma)$ are given by $z = \gamma p_i$. Since multiplication of the complex zero by a scalar only affects its magnitude, we can conclude that the window causes the zeroes of $A(z)$, and consequently the poles of $1/A(z)$, to be shifted into the direction of the origin of the z-plane.

This appendix discusses the relation between the value of γ and the corresponding increase in bandwidth. We start with deriving an expression for the bandwidth of a filter with impulse response

$$x[n] = \gamma^n \upsilon[n], \tag{D.5}$$

where $\upsilon[n]$ represents the discrete-time unit step function. Because this yields an inconvenient expression we also derive an expression for the bandwidth of a continuous-time filter with impulse response

$$h(t) = e^{\frac{t}{T}\ln\gamma}\upsilon(t), \tag{D.6}$$

which equals $x[n]$ at the times nT, where T is the sampling period, and $\upsilon(t)$ is the continuous-time unit step response. We show that the continuous-time bandwidth can very well represent the bandwidth of the discrete-time case.

Bandwidth of the discrete-time γ-window

The z-transform $X(z)$ of the discrete-time impulse response $x[n]$ is given by

$$X(z) = \frac{1}{1-\gamma z^{-1}}. \tag{D.7}$$

Evaluation of the absolute value of this function on the unit circle yields

$$\left|X(e^{j\theta})\right| = \frac{1}{\sqrt{1-2\gamma\cos\theta+\gamma^2}}. \tag{D.8}$$

For $\theta = 0$, the absolute value of this frequency function becomes its maximum value $\frac{1}{1-\gamma}$ and for $\theta = \pm\pi$ it becomes its minimum value $\frac{1}{1+\gamma}$.

If we define the bandwidth as the frequency span between the 3-dB attenuation points with respect to the maximum, then this bandwidth only exists if

$$\frac{1}{\sqrt{2}}\frac{1}{1-\gamma} \geq \frac{1}{1+\gamma}, \tag{D.9}$$

so, if $\gamma \geq 0.1716$. For valid values of γ the bandwidth B is given by

$$B = 2\arccos\left\{1 - \frac{(1-\gamma)^2}{2\gamma}\right\}, \tag{D.10}$$

and the bandwidth b_d in Hertz becomes

$$b_d = \frac{f_s}{\pi}\arccos\left\{1 - \frac{(1-\gamma)^2}{2\gamma}\right\}, \tag{D.11}$$

where $f_s = 1/T$ stands for the sampling frequency.

Bandwidth of the continuous-time γ-window

The Fourier transform $H(\omega)$ of the continuous-time impulse response $h(t)$ is given by

$$H(\omega) = \frac{1}{-\frac{1}{T}\ln\gamma + j\omega}, \tag{D.12}$$

and the absolute value is

$$|H(\omega)| = \frac{1}{\sqrt{(\frac{1}{T}\ln\gamma)^2 + \omega^2}}. \tag{D.13}$$

It is clear that the 3-dB bandwidth of $|H(\omega)|$ amounts to $-2\frac{1}{T}\ln\gamma$. We can also write the bandwidth b_c in Hertz as

$$b_c = -\frac{f_s}{\pi}\ln\gamma. \tag{D.14}$$

Comparison of the two cases

If we compare Equations D.11 and D.14 we see a common factor f_s/π. We will call the remaining factors the normalised expansion bandwidths. These are different in the continuous-time case and the discrete-time case. Figure D.1 shows a plot of these normalised expansion bandwidths. We see that for values of γ close to one the difference between the two bandwidths becomes very small. For values of γ exceeding 0.7 the difference is less then 1 percent. The usually applied values are in the range between 0.7 and 1. So, we can conclude that in this range of γ the discrete-time bandwidth is very well represented by the continuous-time bandwidth.

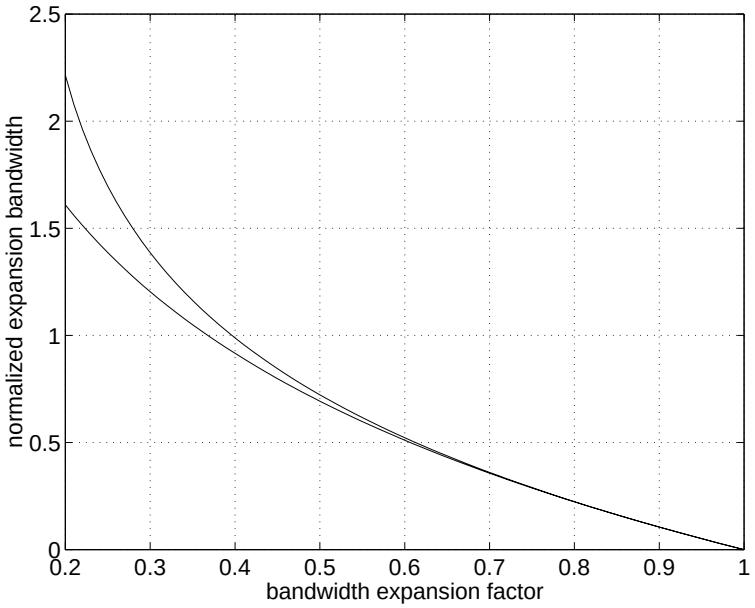

Figure D.1: Comparison of the normalised expansion bandwidths for the discrete-time case (upper trace) and the continuous-time case (lower trace) as a function of the bandwidth expansion factor (γ).

The increase in bandwidth

It has been shown on the basis of Equation D.7 that the bandwidth associated with a real pole with radius $r = \gamma$ in the z-plane is given by Equation D.11. Because of symmetry reasons it is clear that any rotation of the real vector representing such a pole gives rise to a complex vector $re^{j\theta}$ having the same bandwidth. It has also been shown that the bandwidth according to Equation D.11 is very well approximated by Equation D.14, if we assume the radius to be within the range 0.7 to 1. So, the bandwidth b associated with a pole described by $re^{j\theta}$ is very well approximated by

$$b = -\frac{f_s}{\pi} \ln r. \tag{D.15}$$

It has been shown as well - on the basis of Equation D.2 and continuation - that the application of a γ-window shifts a pole into the direction of the origin by a factor of γ. The associated expanded bandwidth b_{ex}

accordingly becomes

$$b_{ex} = -\frac{f_s}{\pi}\ln(\gamma r) = -\frac{f_s}{\pi}\ln r - \frac{f_s}{\pi}\ln\gamma \qquad \text{(D.16)}$$

This expression clearly shows that the bandwidth associated with a pole is increased by $-\frac{f_s}{\pi}\ln\gamma$ when a γ-window is applied.

Example

Figure D.2 shows an example of a 10^{th} order LPC spectrum $1/|A(z)|$ and the bandwidth-expanded spectrum $1/|A(z/\gamma)|$ with $\gamma = 0.95$. The increase in bandwidth according to Equation D.16 amounts to 130 Hz. The sampling frequency is 8 kHz.

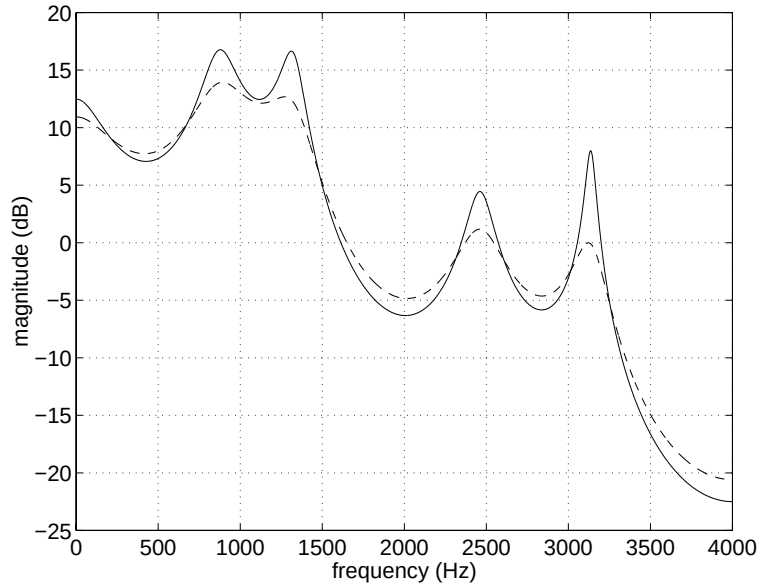

Figure D.2: Example of an LPC magnitude spectrum (solid line) and its bandwidth-expanded version with $\gamma = 0.95$ (dashed line).

Samenvatting

In de kern van dit proefschrift (Hoofdstuk 4) worden de algoritmische achtergronden en de ontwerpaspecten beschreven van de eerste spraakcoder voor publieke mobiele telefonie, de GSM Full-Rate coder, zoals deze gestandaardiseerd werd in 1988. Dit is niet eerder in zoveel detail beschreven. De coder wordt in een historisch perspectief geplaatst door twee voorafgaande hoofdstukken over achtereenvolgens de geschiedenis van spraakproduktiemodellen en de ontwikkeling van spraakcoderingstechnieken, beide tot het midden van diezelfde jaren tachtig. In de epiloog wordt een beknopt overzicht gegeven van de latere ontwikkelingen op het gebied van spraakcodering.

De inleiding, Hoofdstuk 1, begint met enkele uitgangspunten. Er wordt gedefinieerd wat spraakcodering is en de lezer wordt geïntroduceerd in spraakcoderingsstandaarden en de instituten die deze standaarden vaststellen. Dan worden de kenmerken van een spraakcoder die een rol spelen bij het standaardisatieproces beschreven. Vervolgens worden verschillende toepassingen van spraakcoders - inclusief mobiele telefonie - besproken en tenslotte wordt de stand van de techniek in spraakcodering toegelicht aan de hand van enkele wereldwijd erkende standaards.

Hoofdstuk 2 begint met een samenvatting van de eigenschappen van spraaksignalen en van hun bron, het menselijk spraakorgaan. Dan worden historische spraakproduktiemodellen die de basis vormen van verschillende soorten moderne spraakcoders besproken. Beginnende met een overzicht van antieke mechanische modellen komen we gaandeweg uit bij het elektrische bron-filter model van de jaren dertig. Dan volgt een terugblik op de opkomst van de akoestische-buis modellen in de jaren vijftig en zestig. Tenslotte worden de jaren zeventig besproken die het tijddiscrete filtermodel op basis van lineaire predictie brachten. Op unieke wijze worden de logische opeenvolging van deze modellen en hun onderlinge verbanden belicht. Alhoewel de historische modellen worden

behandeld in verteltrant, worden de akoestische-buis modellen en de op spraak toegespitste lineaire predictietechnieken meer onderworpen aan mathematische analyse om een solide basis te creëren voor de verhandelingen van Hoofdstuk 4. Deze trend zet zich op sommige plaatsen in Hoofdstuk 3 voort ter completering van die basis.

In Hoofdstuk 3 wordt de lezer bij de hand genomen op een reis door de tijd waarbij opeenvolgende coderingsmethoden de revue passeren. Op originele wijze wordt speciale aandacht besteed aan het evolutionaire aspect. Voor elke nieuw voorgestelde methode die we beschrijven wordt besproken wat deze toevoegde aan de stand van de techniek van de betreffende tijd. Na de introductie van de relevante voorlopers, beginnende met Pulse Code Modulation (PCM) en de eerste vocoders uit de jaren dertig, arriveren we uiteindelijk bij Residual-Excited Linear-Predictive (RELP) coders, analyse-door-synthese systemen en Regular-Pulse Excitation in 1984. Deze laatste techniek vormt de basis van de GSM Full-Rate coder.

In de kern van dit proefschrift, Hoofdstuk 4, worden expliciete vormen van Multi-Pulse Excited (MPE) and Regular-Pulse Excited (RPE) analyse-door-synthese systemen ontwikkeld. Beginnende bij bekende methoden uit 1984 om de pulsamplitudes te bepalen, die inhielden dat tweehonderd maal per seconde een stelsel vergelijkingen met $10 - 16$ onbekenden moest worden opgelost, worden er vervolgens verschillende expliciete-vorm algoritmes gepresenteerd waarin het oplossen van vergelijkingen in real-time wordt vermeden. Dan wordt het ontwerp van een speciale expliciete-vorm RPE coder en een daaruit voortvloeiende rekenefficiënte codeerarchitectuur beschreven. Deze architectuur maakte het mogelijk om een real-time RPE coder te implementeren op een enkele digitale signaalprocessor van die tijd. De expliciete vormen en de resulterende codeerarchitecturen zijn nergens eerder in zoveel detail beschreven. Vervolgens wordt een implementatie van zo'n coder met een bit rate van 13 kbit/s beschreven die in 1988 geselecteerd werd als de Full-Rate GSM standaard. Tenslotte worden de prestaties van deze standaardcoder gerecapituleerd.

Hoofdstuk 5 is een epiloog met een kort overzicht over de belangrijkste ontwikkelingen in de spraakcodering die plaats hebben gevonden na 1988. Sindsdien zijn er veel nieuwe spraakcoderingsstandaarden gekomen, zowel voor mobiele telefonie als voor andere toepassingen. Het hoodstuk wordt afgesloten met een vooruitblik.

Biography

I was born on July 12, 1946 in Nijmegen, The Netherlands. In 1962 I completed my secondary education MULO A and B. In 1965 I completed an evening training course "Elektronentechniek" in electronics. In 1966 I completed an evening training course "VAKE" (voorbereidend "AKE") at the Instituut voor Hoger Beroeps Onderwijs (IHBO) Eindhoven, which was the name of the Polytechnic ("HTS") at the time. In 1968 I completed the evening training course AKE (Applicatie Kursus Elektronica) at the IHBO Eindhoven. This was a post-HTS-E training course. Next, I have been active as a part-time student at the Technische Universiteit Eindhoven for another period of two years, in order to level up my skills in mathematics, physics and electronics.

I joined Philips Research Laboratories in 1962. Since the late sixties I have been engaged in several research programmes, first as a research engineer and later also in managerial functions as a project leader and as a cluster leader. The programmes concerned data transmission, digital signal processing of speech, audio and video signals - at a mathematical/algorithmic level as well as at the level of VLSI architecture - speech analysis and resynthesis, speech coding for different applications, coding and wireless transmission of audio signals and recently, digital signal processing for the analysis of maternal and fetal ECGs.

Already in the seventies I had a teaching task at a Philips internal course on electronics. I have given internal concern-wide courses on speech coding and occasionally also at the Eindhoven Embedded Systems Institute (EESI). Recently, I teach a course ("keuzevak") on speech coding at the Technische Universiteit Eindhoven.

Since 1993 I am a Research Fellow of the Philips Research Laboratories. I am (co)author of 51 publications and I have given many presentations at conferences and symposia. I am (co)inventor of 25 granted USA patents and I am a Senior Member of the IEEE.