PERCEPTUAL MODELS IN SPEECH QUALITY ASSESSMENT AND CODING

by

Vasos E. Savvides,

BSc, MSc, A.M.I.E.E

A Doctoral Thesis submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of the Loughborough University of Technology

November 1988

SUPERVISOR : Professor Costas S. Xydeas,
MSc, PhD, CEng, MIEE, MIERE, MIOA

ABSTRACT

The ever-increasing demand for good communications/toll quality speech has created a renewed interest into the perceptual impact of rate compression. Two general areas are investigated in this work, namely speech quality assessment and speech coding.

In the field of speech quality assessment, a model is developed which simulates the processing stages of the peripheral auditory system. At the output of the model a "running" auditory sprectrum is obtained. This represents the auditory (spectral) equivalent of any acoustic sound such as speech. Auditory spectra from coded speech segments serve as inputs to a second model. This model simulates the information centre in the brain which performs the speech quality assessment. The second model produces a multidimensional distortion space as its output. Each speech segment is represented by a point (or vector) in this space. The origin of the space is occupied by the uncoded speech segment. Distortion directions (axes) are identified within the space. The projection of the vector corresponding to a speech segment, onto a distortion axis, gives a measure of the amount of that particular distortion contained within the speech segment. Comparisons with subjectively determined distortion spaces confirm the validity of the approach.

Perceptual knowledge is also integrated into the general area of speech coding. Variable rate coding subject to perceptual criteria is investigated. A new speech algorithm is developed. The new algorithm is shown to be a logical development to the line of thought of RELP coders. These employ a LPC vocoder model where the excitation signal is provided by a waveform coded, baseband signal. In the proposed algorithm multiple "basebands" are used whose location and number are made to adapt to the speech signal's short time spectrum. Efficient rate compression algorithms such as subband coding, vector quantization and multipulse LPC are utilized in the new algorithm to produce good communications quality at bit rates around 4.8 Kbits/sec.

Acknowledgments

I would like to express my gratitute to my supervisor, professor C. Xydeas, for his continuous support and encouregement during the course of this research.

The support of the RT5 division of British Telecom Research Laboratories, Martlesham Heath, Ipswitch, in carrying out this work is gratefully acknowledged by the author.

Many thanks also to the speech and image processing research group for their stimulating discussions, help and friendship offered.

Finally, I would wish to thank Sandra for typing this thesis.

Page

CHAPTER 1

INTRODUCTION 1

CHAPTER 2

HEARING 8

2.1 Physiological observations and measurements 9
2.1.1 Anatomy of the Auditory System - An Overview 9
2.1.3 The inner ear 10
2.1.3.1 The Basilar Membrane 10
2.1.3.2 The outer and inner hair cells 14
2.1.3.3a Intensity and frequency effects 16
2.1.3.3b Gross time behaviour 19
2.1.3.3c Detailed time behaviour 19
2.1.3.3d Responses to impulses 23
2.1.3.3e Speech coding in the auditory nerve 24

2.2 Models of physiological characteristics of the Peripheral Auditory System 42
2.2.1 Models of hydrodynamics 42
2.2.2 Models incorporating mechanical to neural transduction mechanisms. 43

2.3 Psychophysical observations and measurements 51
2.3.1 Loudness and masking effects of simple sounds 51
2.3.1.1 Loudness level 51
2.3.1.2 Loudness 51
2.3.1.3 Masking effects 52
2.3.1.4 Loudness curve near threshold 53
2.3.1.5 Masking of pure tones by pure tones. 55
2.3.2 Psychoacoustical tuning curves - the critical band 56
2.3.3 Partial masking - Asymmetry between noise and tone maskers 65
2.3.4 Temporal effects 72
2.3.4.1 Temporal summation 72
2.3.4.2 Transient effects in masking: Backward M., Forward M. and Overshoot. 74
2.3.4.3 Transient effects in Masking: Broad band noise masker 77
2.3.4.3a Simultaneous Masking 77
2.3.4.3b Forward Masking 78
2.3.4.3c Backward Masking 80
2.3.4.3d Transient Masking patterns 81

2.3.4.4 Transient effects in masking. Critical band noise masker. 82
2.3.4.4a Simultaneous masking 83
2.3.4.4b Forward masking 83
2.3.4.4c Backward masking 84
2.3.4.4d Transient masking patterns 84
2.3.4.5 Transient masking effects: Sinusoidal masker 85
2.3.4.6 Statistical changes 86
2.3.4.7 Masking period patterns 90
2.3.5 Discrimination (difference limens, just noticeable difference) 99
2.3.5.1 Intensity discrimination 100
2.3.5.2 Frequency discrimination 101
2.3.5.3 Discrimination of Complex tones 101
2.3.6 Phase Perception 104
2.3.7 Roughness 109
2.3.8 Pitch 112
2.3.9 Timbre 116

2.4 Models of Psychoacoustical phenomena 144
2.4.1 The Critical band scale 145
2.4.2 A model of loudness summation 146
2.4.3 The excitation pattern in other phenomena 153
2.4.3.1 Difference limens 153
2.4.3.2 Pitch and phase perception 155
2.4.3.3 Roughness 155
2.4.4 Modelling the time effects - an improved model 155
2.4.5 The model by Schroeder *et al* 159

CHAPTER 3

SPEECH AND SPEECH CODING 176

3.1 Speech: Production and Phonetics 176
3.1.1 Acoustical speech production 176
3.1.2 Phonemes 177
3.1.3 Acoustic-phonetics 178
3.1.4 Modelling the speech production process 179

3.2 Speech coding 183
3.2.1 Introuction 183
3.2.2 Vocoders 186
3.2.2.1 Channel vocoders 186
3.2.2.2 LPC vocoders 187
3.2.3 Waveform coding 191
3.2.3.1 Pulse Code Modulation (P.C.M.) 191
3.2.3.1a Non-uniform quantization 194
3.2.3.1b Logarithmic quantication 195
3.2.3.1c Entropy coding 197
3.2.3.1d Representation of Noise 198

3.2.3.1e Adaptive quantization 199

3.2.3.2 Differential P.C.M. (D.P.C.M.) 202

3.2.3.2a The linear predictor of order P 206

3.2.3.2b Adaptive prediction 208

3.2.3.2c Adaptive predictor design 210

3.2.3.2c.1 The auto-correlation method 210

3.2.3.2c.2 The co-variance method 213

3.2.3.2c.3 Lattice formulations - The Burg method 216

3.2.3.2d Split band LPC 218

3.2.3.2e LPC quantization 222

3.2.3.2e.1 Optimum bit allocation 224

3.2.3.3 Vector quantization (VQ) 226

3.2.3.3a Codebook generation - the LBG algorithm 229

3.2.3.3b Applications - nearest neighbour measures 231

3.2.3.3c Other issues in vector quantization 238

3.2.3.4 Backward predictor adaptation 241

3.2.3.5 Spectral fine structure predictors 242

3.2.3.6 Noise feedback coding - Noise shaping 247

3.2.3.7 Delayed Decision Coding 253

3.2.3.8 Frequency Domain Coders 256

3.2.3.8a Subband Coding 257

3.2.3.8a.1 Integer band filter banks 260

3.2.3.8a.2 Quadrature Mirror filters 261

3.2.3.8a.3 Delay 265

3.2.3.8a.4 Frequency Inversion 265

3.2.3.8a.5 Transmission rate, S.N.R. and Gain over P.C.M. 266

3.2.3.8a.6 Speech Coding in Subbands 270

3.2.3.8b Transform coding 273

3.2.3.9 Hybrid coding 277

3.2.3.9a Voice Excited Vocoders (V.E.V.) 277

3.2.3.9b Residual Excited Linear Prediction (R.E.L.P.) coders 278

CHAPTER 4

DISTORTION MEASUREMENTS 317

4.1 Rate Distortion Theory 317

4.2 Objective Distortion measures 321
4.2.1 Spectral distance measures 326
4.2.2 Perceptually Motivated measures 330
4.2.3 Modelling the auditory mechanism 331
4.2.4 Separating the distortion from the signal 338
4.2.5 Measuring the distortion 339
4.2.6 The complete models 343
4.2.7 Objective and subjective tests 344
4.2.7.1 Test material 344
4.2.7.2 Subjective tests 345
4.2.7.3 Objective tests 345
4.2.8 Multidimensional representation 352
4.2.9 Applications to coding - selection of threshold parameters 357

4.3 Subjective measurements of coder quality 358

CHAPTER 5

VARIABLE RATE CODING 383

5.1 Introduction 384

5.2 Variable rate coding subject to a perceptual criterion 387

5.2.1 "Real-time" algorithms (suitable for speech transmission) 387
5.2.2 "Non-real-time" algorithms (suitable for speech storage) 400

5.3 Variable rate coding based upon a w.m.s.e. criterion 403

CHAPTER 6

A NEW APPROACH TO LOW BIT RATE SPEECH CODING 432

6.1 Introduction 433

6.2 Side Information: LPC filters 435

6.3 Side Information: Gains 446

6.4 Bit allocation 451

6.5 Pulse allocation 456

6.6 Gains coding 458

6.7 Pulse Amplitude Coding 462

6.8 Preferred Simulation Parameters for a 4.8kbits/s coder 464

6.9 Speech quality 466

6.10 An improved algorithm 467
- 6.10.1 Choice of passbands 470
- 6.10.2 Pulse allocation 472
- 6.10.3 Pulse allocations and amplitudes 475
- 6.10.4 Stopband regeneration 478
- 6.10.5 A modification to the coding of the LPC parameters 480
- 6.10.6 Speech quality obtained from the final algorithm 482
- 6.10.7 Coder complexity 483
- 6.10.8 Further work 483
- 6.10.9 Note on Publication 486

CHAPTER 7

RECAPITULATION AND CONCLUSIONS - DIRECTIONS FOR FURTHER RESEARCH 500

Appendix A: Sound intensity measurements 512

Appendix B: octave band filter and derivatives 514

Appendix C: Positive integer constrains on the bit allocation formula 516

Appendix D: Solution to the generalized bit allocation formula 522

Appendix E: Prediction gain from Rate Distortion Theory 531

Appendix F: The Multipulse Algorithm 533

Appendix G: Pulse Allocation Procedure 542

Appendix H: A comparative study of coder complexity. 545

CHAPTER 1:

INTRODUCTION

CHAPTER 1: INTRODUCTION

In recent years and particularly throughout the 1980's a tremendous amount of research and development has been carried out internationally in the field of Digital Speech Coding.

New applications areas such as the Integrated Services Digital Network (ISDN) and digital mobile radio provided the resources to fuel research in this field.Nevertheless, the impetus has been provided by the excellent Digital Signal Processing (DSP) chips that have been made available to rate compression workers. Particular emphasis has been placed upon low bit rates in the range of 4.8 to 9.6 kbits/sec with the objective of providing Toll to good Communications quality speech: The cost of a Telecommunications Network is the sum of the source compression cost plus the actual cost of providing the transmission path. The increased availability of DSP chips has resulted in a reduction in the source compression cost. In response to this reduction, organizations involved in the provision of Telecommunications services strived to push down the transmission rate (whilst maintaining "acceptable" quality speech) in order to utilize the available bandwidth as profitably as possible.

As bit rate is reduced the minimum achievable reconstruction noise increases.It is therefore important at these rates to operate as close to rate distortion bounds as possible,through the use of complex and sophisticated algorithms. The newly developed Delayed Decision Coding Algorithms such as Code Excited Linear Predictive Coder (CELP) and the Multipulse Linear Predictive Coder attempt to fulfil this objective.

Also of paramount importance, is an understanding of the concept of a perceptually meaningful distortion criterion. The ultimate test of each and every coder structure is its subjective quality. Speech coders should therefore be optimized and assessed through the use of subjective tests. Unfortunately, such tests are time consuming and expensive to conduct. Furthermore, their outcome usually provides limited information, relevant only to the particular conditions and coders used in each test. General conclusions cannot usually be drawn (or the results are not very useful) for optimizing those coders under test or others. Although the use of multidimensional diagnostic tests, such as the ones carried out and presented in the latter part of chapter 4, offer some improvement toward this situation it is only through the use of perceptually meaningful objective tests that rapid growth and improvement can be expected.

In order to be able to construct a useful objective measure, capable of postdiction and prediction of subjective tests and results, it is necessary to have a deep knowledge and understanding of the mechanisms of auditory perception. Two distinct aspects of auditory perception can be identified: One is related to the peripheral auditory processing which is carried out within the ear itself up to the auditory nerve.This constitutes a front end processor to the second part which resides in the higher auditory pathways and in the brain. This latter part forms the Information processing and cognition centre of perception; Very little is known about the second part, traditionally of interest to social scientists (e.g. psychologists), but now becoming the subject of Information Technology. Due to its peripheral nature, it has been possible to investigate auditory processing to a greater degree. Even so, the peripheral auditory processing part is still a very dark area compared to other fields of interest to speech coding specialists, for example, speech production. This is undoubtedly due to the severe nonlinearties of the system which make it very difficult to model. One additional limitation is the complexity of isolating the effects of the peripheral front end part from those of the central part in any relevant experiment.

Knowledge about the peripheral auditory system (PAS) comes from two different research methodologies: The first is physiological and concerns direct physical and physiological measurements. These are usually obtained through surgery and subsequent electrical recording from the auditory nerve fibres. These carry the auditory information from the ear to the brain in electrical form. Due to their nature, nearly all experiments of this type (with perhaps the only exception being that of cochlear implants) involve non-human subjects. In both types of experiments (physiological as well as psychophysical) the procedures used are almost of equal importance to the results obtained, since it is very difficult to draw any conclusions outside the context of the experiments themselves. In the first part of chapter two a comprehensive range of physiological phenomena and measurements are described. These relate to the motion of the hydromechanical parts and in particular the Basilar Membrane, the mechanical to neural transduction in the hair cells and finally, to the nerve fibres which convey the information to the brain. The emphasis is on the response of each part to various stimuli and its aim is to show how each part affects the shape and transforms the nature of the acoustic waveform signal as it passes successively through the different stages to arrive, in electrical form,at the nerve fibres on its way to the brain. In order to describe the system with any degree of accuracy, and in the absence of any linear behaviour, it becomes necessary to provide the response of the system to a broad selection of stimuli.

The first part of chapter two is supplemented with a small section on various models that have been devised to simulate physiological behaviour.

Having obtained a certain knowledge about the physiological response of the PAS one is now able to understand and utilize the other form of knowledge that is available about the PAS, which derives from psychophysical experiments. Psychophysical experiments involve one or more physical stimuli and the corresponding response from the subject. This is usually a rating of the impact of the stimuli according to the subject's own psychophysical scales. The averaging of data among subjects (were permissible) provides information on the response of the "average" observer. The last half of chapter two deals with a range of psychophysical phenomena relevant to speech quality evaluation. These include the more conventional Loudness and Masking effects of constant or time independent stimuli as well as the equivalent phenomena for temporally variable sounds. The emphasis is on the evolution of auditory representations with time as opposed to the traditional "static" representations, so that the information is relevant to speech signals which have a high information content and thus vary both in time and in frequency.

The contents of chapter 2 derive from a selection of journals and conference publications from a broad range of fields including physiology, psychophysics, phonetics, acoustics, as well as engineering. It intends to serve both as a necessary review and introduction to the rest of the thesis as well as a reference source for those who need to acquire a broad knowledge and understanding of auditory perception as applied to speech coding.

Chapter 3 starts with a brief discussion of speech production and speech phonetics. This aims to describe the type of signal that one is faced with in the speech processing domain. The main part of chapter 3 is devoted to speech coding. Since several good texts now exist on speech coding only those aspects of speech coding used in subsequent chapters are to be found here. For the purposes of completeness there is a brief mention of other areas whilst certain general structures such as Delta Modulation are not mentioned at all. Nevertheless the author's objective is to present the subject through a logical progression and ascent in algorithm complexity, as soon as the basic concepts are covered. In addition, results are presented with a perceptual perspective in mind, whenever possible.

Chapter 4 is a treaty on distortion measures. Rate distortion theory serves as an introduction and a fairly comprehensive list of distortion measures that have appeared in the literature is given, together with the extent of applicability of each measure. The need for a perceptually motivated objective measure is justified and various such measures are developed and tested. These are based firmly on the theory presented both in chapter 2 and chapter 3. Towards the end of the chapter multidimensional scaling is used to decompose and diagnose perceptually distinct types of distortion as measured in both objective and subjective tests. Procedures for obtaining such subjective tests are presented.

Chapter 5 deals with variable rate coding subject to perceptual criteria. Algorithms for both "real-time" and "background task" type applications are developed and tested.

Chapter 6 presents a new speech coding structure which combines perceptual knowledge with efficient rate compression algorithms. Although one particular implementation is presented here, the concept of the coder itself is more general and provides a new, generalized approach to speech coding. These ideas are expanded more in the last chapter, chapter 7 which provides the conclusions to the present work and indicates the research areas which the author considers worthy of more investigation and research. These are related to most of the work covered in this thesis.

Finally, the first two appendices, A and B, relate to sound and sound Intensity measurements and are a necessary prerequisite to the reading of chapter 2. The rest of the appendices are related to specific parts within the rest of the thesis and are not intended for general reading. The reader is advised to turn to these appendices when particular mention is made in the relevant part in the thesis, to clarify the particular part in question, or satisfy himself of a proof to a particular equation.It should be noted that most of the appendices cover original work by the author. This is the case, for example, for the bit allocation procedures described in appendices C and D.

CHAPTER 2 HEARING

2.1 PHYSIOLOGICAL OBSERVATIONS AND MEASUREMENTS

2.1.1 Anatomy of the Auditory System - An Overview

A pictorial description of the ear is shown in figure 2.1-1a. It is generally divided into three regions, the outer, middle and inner ear, as shown in figure 2.1.1c

Sound is transmitted through the external auditory meatus (the ear canal) and causes the tympanic membrane to vibrate. This vibration is transmitted *via* the ossicles (figure 2.1.1b) through the oval window to the fluids which fill the cochlea (figures 2.1-1c and figure 2.1-2).

The movement of the cochlea fluids cause travelling waves to be formed on a membrane running along the length of the cochlea, the Basilar membrane (BM) (figure 2.1-2). The waves originate from the stapes and travel round the spirals of the cochlea towards its end at the helicotrema (figure 2.1-3).

The travelling waves bend the outer and inner hair cells (figure 2.1-2c) which constitute the mechanical to neural transducers of the peripheral Auditory system (PAS).

The (electrical) signals generated there are transmitted through the nerve fibres connected to them, to higher auditory pathways. These fibres constitute the auditory nerve.

2.1.2 The outer and middle ear

The outer ear enhances the acoustic signal in the 2000 to 6000 Hz region. This enhancement is of the order of 10 dB[a] and is one factor that contributes to the ears maximal sensitivity to frequency signals in this region. This can be seen in a plot of the threshold of hearing versus frequency (figure 2.1-4).

The middle ear acts as an impedance transformer. It reduces the impedance mismatch between the air outside the ear drum and the cochlea fluids. Together with the outer ear they behave as a lowpass filter with an "overshoot" in the 2 to 5 KHz region reaching about 8 dB at 3.5 KHz [c]

2.1.3 The inner ear

2.1.3.1 The Basilar membrane:

When the oval window is set in motion by an incoming sound, a pressure is applied by the fluids of the cochlea essentially simultaneously along the whole length of the BM. This is due to the high speed of sound in the cochlear fluids [160,000 cm/sec, hence, for a 10 KHz sinosoid the wavelength is 16 cm], and the small size of the cochlear duct (about 3.5 cm in man).

The response of the BM to sinosoidal stimulation begins as a bulge at the basal end (near the stapes). The bulge takes the form of a travelling wave which moves along the basilar membrane towards the apex (near the helicotrema). The amplitude of the wave increases slowly at first and then decreases rather abruptly. This is shown in figure 2.1-5. The position of the peak in the pattern of vibration varies according to the frequency of stimulation. High frequency sounds produce a maximum displacement of the basilar membrane near the basal end so that there is little activity on the remainder of the membrane. Low frequency sounds produce a pattern of vibration which extends all the way along the BM but which reaches a maximum before the end of the membrane. Figure 2.1-6 shows the envelopes of the vibration patterns for several different low frequency sinosoids.

From figure 2.1-6 it can be seen that the mode of vibration to a low frequency sinosoid of sufficient amplitude can "swamp" the vibration produced by a high frequency sinosoid of smaller amplitude ([d] p.93). In response to steady sinosoidal stimulation each point on the basilar membrane vibrates in a sinosoidal manner with a frequency equal to that of the input waveform ([d] p18).

Several workers have produced tuning curves for the Basilar membrane (figure 2.1-7). It should be noted that the results shown have been obtained using different methods and also different species by each worker.

Most of the pioneering work on vibration patterns along the BM was done by Von Békézy. These early experiments revealed rather broad tuning curves for the equivalent BM "Filters". This presented a problem for auditory theorists since the observed frequency selecticity was insufficient to explain various psychoacoustical phenomena which implied a much better frequency resolving power. The methods used by Békézy suffered from certain shortcomings: Békézy used human cadaver ears. Later it was found that soon after death the mechanical properties of the BM change significantly. [5]. Also, the vibration amplitudes had to be at least of the order of one wavelength of visible light. This required very high sound levels (about 140 dB SPL). If the response of the BM at these levels was non-linear, the tuning curves might have been much sharper at lower (normal) levels. Recent techniques do not have to rely on direct visual observation of the BM. For example, the Mössbauer technique measures doppler shift in the emmitted radiation of gamma rays, from a very small radioactive source placed on the BM. At high frequencies (~7KHz) normal SPL levels (70-80 dB) can be used to produce reasonable results. In these experiments live animals were used, with auditory systems similar to that of man (guinea pig and squirrel monkey).

Rhode [4] produced both amplitude and phase curves for the BM, versus input frequency. All his measurements were made at points maximally sensitive to about 7-8 KHz.

From the amplitude versus frequency curves he deduced that the low frequency slope of the amplitude curve was 6 dB/octave. As frequency was increased and the maximum amplitude was approached the slope increased to 24 dB/octave. The slope of the curve beyond the maximum amplitude was about 100 dB/octave. Also, near the frequency of maximum displacement (characteristic frequency) he found an amplitude nonlinearity.

From the phase plots he deduced that at low frequencies (100-400 Hz) the phase was $\pi/2$. For these frequencies the whole BM vibrates in phase.

For higher characteristic frequencies the phase is a linear function of frequency over most of the lower frequency range. Near the cutoff region the phase increases at a faster rate.

For frequencies greater than the frequency corresponding to the maximum amplitude response the phase approaches a constant value.

Another important aspect closely related to the phase response is the velocity of propagation of the travelling wave. Rhode's results show that the velocity of propogation of the travelling waves is approximately constant for frequencies up to the maximum amplitude point. Above the maximum the velocity decreases slightly but then, at very high frequencies, it increases dramatically (the range of velocities before the maximally effective frequency was from 12-9 m/sec). Note that all the above effects refer to only a small region on the BM, and that the independent variable here was the input frequency and not the location on the BM. Typical curves from Rhode are shown in (figure 2.1-8).

From indirect measurements Anderson *et al* [2] deduced that a linear relationship (on the average) existed between frequency and phase response. This relationship was used to find a time delay to each point on the BM (which was independent of frequency). They found that the travel time decreased exponentially with increasing distance from the oval window (and was independent of input frequency).

The main implication of travel time being independent of frequency (or equivalently a linear phase versus frequency plot) is that the waveform of a complex stimulus tends to be preserved when mechanically propagated along the cochlear partition.

Dallos [3], also from indirect measurements, found that the travelling wave velocity was frequency independent. They calculated that the travel time for a point 10 mm from the stapes was 0.62 msec, whereas, for a point 14 mm away the time was 1.4 msec (for a guinea pig cochlea). For the former value this was true for a range of frequencies form 100 Hz to 5 KHz.

Several workers seem to disagree with the linear phase versus frequency relationship [7,6] although their views are largely dependent on Békézy's experiments [1]. These were made at low frequencies where the helicotrema appears to have some effect [9].

With respect to variations with distance, the velocity of propagation decreases exponentially with distance from the stapes (as does the travelling time).

Also the wavelength of the travelling wave decreases monotonically with distance [11].

The above experiments although performed using simple stimuli (sinosoids) indicate the fashion in which the BM can be represented, as a spectrum analyser, with finite frequency resolution; Components of a sound, sufficiently far apart in frequency, will be resolved without too much interaction. For components relatively close in frequency, however, the patterns of vibration of the BM will interact and when the components are sufficiently close in frequency, then, there will no longer be a separate maximum for each component in the vibration but, instead, a single broader maximum.

Above 500 Hz the resolution of the analyser is proportional to centre frequency. This has important implications for harmonic complex sounds such as the ones occurring in speech and music. Since the spacing of the harmonics is constant the lower frequency harmonics are much better resolved than the higher harmonics. This is shown schematically in figure [2.1-9]. Note that the periodicity of the sound can either be determined from the place of occurrence of the first harmonic or, alternatively, from the time patterns at higher frequency ranges. Therefore the broader response at higher frequencies does not necessarily imply a loss in information since the time waveform within these broad filters may be reserved.

2.1.3.2 The outer and inner hair cells:

These receptors are arranged in an orderly fashion along the extent of the BM. They form part of the organ of cordi. There are about 25000 outer hair cells arranged in three rows with about 140 hairs protruding from each. The inner hair cells are fewer, about 3500, in a single row, and have about 40 hairs each.

The motion of the BM is thought to cause these cells to bend ([e] p.28) and release a signal (series of spikes) picked up by fibres in the auditory nerve. It is generally accepted that the firing of a hair cell depends on the motion of the BM and it is not frequency specific itself. Then, on a simple threshold rule, the tuning curve for a cell should be the mirror image of the tuning curve for the BM ([e] p.57) figure (2.1-10) (compare with figures 2.1-6 and 2.1-12).

Studies on the hair cell tuning curves ([d] p.25) indicate that they are more sharply tuned than the patterns of displacement of the BM. This seems to support views that the firing probability of a hair cell is proportional to some derivative of the BM motion ([f] p.25).

It's not exactly known how this transduction operates. Some results indicate that the response of the outer cells is proportional to displacement whereas the response of the inner hair cells is proportional to velocity [12]. Also the bending of the hair cells is produced by longitudinal and radial shear forces (the latter arising from the BM being fixed along both sides of its length, figure [2.1-11]. It is possible that the inner or the outer cells may be less sensitive (or not at all) to one or the other shear forces [9]. Several workers tried to explain additional sharpening of the "tuning curves" from these modes of stimulation of the hair cells and form interactions between outer and inner hair cells. For example, tuning curves due to the shearing displacement appear to have steeper slopes than those from the travelling wave [11,12]. Also, sharpening of the "filters" slopes could result from the directional sensitivity of the cells [14] or the different phase response of the inner and outer hair cells [13].

2.1.3.3 Nerve fibres in the auditory nerve.

There have been extensive studies upon the response of single fibres to various stimuli. The responses are functions of all of the stimulus attributes, namely, the variation in amplitude in both the frequency and time domains. The fibres also show spontaneous activity in the absence of any sound stimulation. Values of 10 to 50 spikes/sec are typical.

2.1.3.3a Intensity and Frequency Effects

First, regarding the dependence of the response upon frequency-intensity variations of the stimulus, the fibres show considerable frequency selectivity. This can be seen in tuning curves in figure [2.1-12]. These represent a cell's threshold as a function of frequency. The threshold is defined as the lower level of a sinosoid for which the experimenter by auditory and visual monitoring of the activity of the fibre can detect a change in the activity of the fibre. Stimuli are usually tone bursts so that changes can be detected more easily. The frequency at which the threshold is at its lowest is called the characteristic frequency (CF).

A single fiber seems to derive its output from a particular part of the BM. Also, fibres with high CF are found in the periphery of the auditory nerve bundle with an orderly decrease in CF towards the centre of the bundle [15].

To describe the characteristics of single fibres at levels above threshold either isorate or iso-intensity contours can be plotted. The former are plots of the intensity of sinosoidal stimulation required to produce a predetermined firing rate as a function of frequency. They are generally similar in shape to tuning curves.

Iso-intensity contours show firing rates at equal sound levels as a function of stimulus frequency. These are very different from the tuning curves [16, 20] (figure 2.1-13). A broadening of the frequency response can be seen at high intensity values. This in a sense reflects the relationship between intensity and firing rates.

The effect of intensity can further be seen in figure [2.1-14]. The figure shows both the spontaneous firing rate and the nonlinear relationship between firing rate and intensity which leads to saturation within 30-40 dB of the input stimulus [17].

It appears that the majority of fibres reach saturation within 40 dB of threshold whereas the threshold values have a range of about 20 dB amongst fibres [15, 17]. Hence only an intensity range of about 60 dB seems to be coded in the fibers. Recently though fibres with much higher thresholds have been reported [18] although they only represent a small percentage of the total. Another small number of fibres seem to show a change in slope of firing rates versus intensity instead of a saturation effect, thus increasing their dynamic range to about 60-70 dB. [19] Also at higher stimulus intensities neighbouring fibres with close CF values may start to fire (as suggested by the iso-intensity contours, figure 2.1-13) which may be another alternative to coding intensity at higher stimulus levels.

Comparisons with the tuning curves from the BM [21-24] revealed that the curves from nerve fibres are more sharply tuned than the equivalent ones from the BM measurements. This fact led to speculations about a possible "second filter" between the BM and the nerve fibres. One line of thought places this "second filter" within the hair cells as mentioned in the previous section.

As a first approximation, it seems that the 30,000 or so fibres in the mammalian cochlear nerve can be considered functionally to represent a bank of (heavily) overlapping narrow (~ 1/3 octave) bandpass linear filters, each followed by a nonlinear probabilistic, analogue to rate converter. The filtering process seems to be due to a two stage system where the BM provides an initial low-pass response, followed by a second stage of a narrow bandpass filter located between the BM and the cochlear nerve, whose exact location is as yet undiscovered.

Direct measurements of the filters bandwidths seem to correspond (to a first approximation) to the critical bands (in later sections) in man. These constitute a psychophysical measure of frequency selectivity (or resolution) for simultaneously analyzed components of a complex sound. This is not to be confused with the ability for frequency discrimination for sequentially presented frequencies (jnd) which is approximately 30 times finer and probably related to the steep cutoffs of the "filters" instead of their bandwidths. There have been some recent studies on the response of many single neurons to a limited set of stimuli [25-26]. At low levels the response is a high level of activity from fibres with a CF close to the stimulus with activity dropping off on either side of this as predicted in [e] p65-68 (fig. 2.1-15,2.1-16). At high stimulus levels the effects of saturation come into place producing a plateau of uniform high level of activity over a wide range of CFs on both sides of the CF corresponding to the stimulus frequency, with activity falling off at CFs far removed from the stimulus frequency. This again agrees with speculations made by Whitfield [e] (figure 2.1-16).

The distribution of neural activity as a function of CF (over all the fibres) is called the "excitation pattern". It represents the effective amount of excitation produced by a stimulus as a function of CF, (shown schematically in figure 2.1-17). The determination of these excitation patterns for different stimuli is of great importance to loudness calculations and it has become accepted as an internal representation of the spectrum of the stimulus. This is the last stage of the PAS processing accessible to physiological measurements.

The time domain response of the auditory nerve fibres has also been studied extensively. Some results are given below.

2.1.3.3b Gross time behaviour

The Gross time behaviour is examined through the plot of post stimulus time (PST) histograms. To determine a PST histogram the sound is presented many times and the numbers of neural impulses occurring at various times through the course of the sound (and its repetitions) are counted. Zero time is reset each time the sound is applied. These are plotted in the form of a Histogram [15]. Figure (2.1-18a,b) shows the typical time course of the discharge of cochlear fibres to short duration tone (A) and noise (B) stimuli. The rate of firing reaches a maximum within a few msec of stimulus onset and adapts at an increasingly slower rate with time. At the end of the stimulus the firing rate drops below the average spontaneous rate. This behaviour is characteristic of all the fibres and is relatively independent of the nature and parameters of the stimulus i.e. tone, noise, frequency and intensity. The details of the exact shape of the response are level and duration dependent [27].

2.1.3.3c Detailed time behaviour

One way to observe the fine time behaviour of the fibres is through an interval histogram. This is a histogram of the distribution of times between successive neural spikes: The time axis is divided into a number of bins and the number of spikes with interarrival times falling into each bin is accumulated and plotted. Interval histograms for the same fibre taken at several different stimulus frequencies are shown in figure [2.1-19] [28].

It can be seen that the discharges are locked to the cycles of the input waveform and occur at intervals which group around the integral multibles of the period of the stimulus frequency. The time structure of the discharges is largely independent of the best frequency of the fibre or the SPL of the stimulus as long as the stimulus is above threshold, [30]. Although the spikes seem to be locked to a cycle, not every cycle is an effective stimulus. Locking occurs with stimulus frequency up to 3000 HZ but histograms become increasingly more blurred above 2500 Hz. This highest frequency is important in order to determine whether "place" or "time" theories of hearing are correct (in a subsequent section). It is not clear though whether the low pass effect is due to injuries of the fibres during measurements. In addition there is a certain jitter associated with the timing of each fibre which obscures measurements at higher frequencies.

Phase locking is what could be expected to occur as a result of the transduction process: when the BM moves upwards, towards the tectorial membrane, the hair cells are bend and a neural response is initiated. No response will occur when the BM moves downwards. Thus nerve firings tend to occur on a positive deflection of the BM produced by the rarefaction phase of the stimulus. From the interval histograms it seems that the probability of a fibre firing on any period of the stimulus is some fixed value P and that successive periods of the stimulus can be treated as independent events at least for low frequency (<2000 Hz) stimuli. Although the phase-locking of fibres to a single tone is impressive it can only be useful to the auditory system if something analogous occurs for complex sounds as well.

When two pure tones not harmonically related are sounded together and no constant phase relation exists between them then the responses can be locked on the primary or secondary or both tones. Which mode prevails depends on their relative strength [31]. This interaction indicates that the activity of one fibre in response to one tone can be suppressed by the presence of a second tone. This has been called two-tone inhibition (suppression). The phenomenon is investigated by presenting a tone at or near the CF of a fibre. A second tone is then added to the stimulus. The suppressing tone has its greater effect when its frequency is just above or below the fibre's tuning curve. [figure 2.1-20].

More controlled conditions can be created when harmonically related tones phase locked to each other are used. Then, a stable complex periodic waveform is generated which can be systematically affected by a change in amplitude or phase of the component frequencies [29]. In this case the discharges can be related to the period of the complex sound by constructing histograms modulo this period: The zero of the time axis is reset [as in PSTs] once every period in synchrony with the stimulating waveform, figure [2.1-21]. Phase-locking is also evident here. Over a wide range of input intensities (30-100 dB) the distribution of spike activity resembles very closely the amplitude distribution imposed by the stimulus.

It seems that the transduction does not operate on a fixed amplitude threshold. This would be inconsistent with the fact that the waveform is portrayed so accurately over such a wide range of SPLs. It appears that the fibre's response follows the normalised amplitude of the stimulating waveform. This is already evident in period histograms for single tones when the SPL is varied. Since no shifts are shown in the peak response [28] within the period of the stimulus, the mechanism is unlikely to be an amplitude threshold device. The data point to a phase sensitive device with a probability of firing monotonically increasing with the positive (or negative) phase of the stimulus but with the variation in amplitude as a function of time within a period, normalized by the peak response of the vibration.

It can also be deduced that there is no direct relation between the total rate of firing and the stimulus intensity. In certain cases turning on the stimulus will actually suppress the total firing rate below the spontaneous rate especially for fibres with high spontaneous rates.

The phase of the stimulus at which the probability of discharge is maximum differs systematically from fibre to fibre according to the CF, and, for a single fibre, according to the stimulus frequency [32, 2]. In [32] Pfeiffer and Molnar computed the phase lag of the fundamental component from Fourier analysis of period histograms obtained from cat cochlear fibres. For fibres of CF lower than 2 KHz this was an approximately linear function of frequency although for certain fibres a better fit to the data was obtained by two straight lines intersecting near the CF [figure 2.1-22]. From more limited data linear phases versus frequency were obtained [2] from which a total time delay was derived to each fibre. This enabled Anderson *et al.* to obtain estimated "travel times" of the cochlear disturbance to the points of innervation of the BM. The delay times ranged from 0.3 msec for CFs around 10 KHz to about 5 msec for fibres with CFs of about 200 Hz. These calculations didn't take into account the "response time" of the cochlear filter [33].

2.1.3.3d Responses to Impulses

To study the fibres's response to clicks (impulses), pulses of approximately 100 µsec duration are applied. The responses are analysed with a PST Histogram. Two quantities can be derived from the histograms. The first is the latency from the onset of the click until the occurrence of the first neural activity. The latency is longest for the low frequency fibres and shortest for the high frequency fibres. A series of pulses can be seen at regular intervals in the PST. The interpeak intervals appear to be related to the characteristic frequency of the fibre.

The latency is again related to the finite velocity of the travelling wave from stapes to the helicotrema in response to the impulsive stimulus. The first impulse reflects the time of arrival of activity at that position in the cochlea. Two or three msec is typical for low frequency fibres (the disturbance has to travel the entire extent of the membrane). The successive spikes after the first reflect the sharp tuning of the fibres. In effect, frequencies very near the CF of the fibre have a dominating effect in the response which resembles the response of the fibre to a sinosoid of frequency equal to the CF of the fibre. Interspike intervals are proportional to the reciprocal of the characteristic frequency of the fibre [15].

The polarity of the click affects the histogram. The click could either be an outward excursion of the earphone, (a condensation click) or an inward motion of the earphone (a rarefaction) click. The histogram obtained with a condensation click roughly equals that obtained with a rarefaction click except that the times are shifted by half a period (half of the interspike interval). This fact once again shows that only one half-cycle of the relative motion of the BM excites a fibre. A "compound" histogram can be obtained by inverting the histogram obtained with the condensation click and adding it to the rarefaction histogram [34]. This histogram now resembles the pattern of vibration caused by the click stimulus [figure 2.1-23].

2.1.3.3e Speech coding in the auditory nerve

In [40, 41], observations on four formant steady state vowels suggested that the profile of average discharge rate across the tonotopically arranged (i.e. arranged according to the CFs) array of auditory nerve-fibres was a poor candidate for representing formant frequencies because virtually all the fibres would be discharging at high rates, for stimulus levels well within those normally used in conversation. Although the debate as to whether average discharge rates are useful for vowel coding is still going on, more recent studies [35-39] have concentrated on how formant pattern and fundamental frequency could be represented in the fine time patterns of discharge of the most sensitive auditory-nerve fibres. (Fibres with high thresholds were discarted). The animal used in the experiments was the cat. Two formant steady state vowels were used as stimuli at a constant fundamental frequency (pitch). The locations of the formants were obtained in perceptual matching experiments with natural vowels. The stimuli were repeated at the rate of 100/min to obtain post-stimulus time (PST) histograms. The bin width was 0.05 msec (which gives a frequency resolution for up to 4kHz). There were 250 to 500 presentations of the stimulus per histogram. Period histograms were computed by adding histogram waveforms for each period. The DFT's of the period histograms were used to estimate the synchronisation indices for each harmonic of the fundamental up to 5 kHz.

The synchronisation index which varies between zero and one indicates how well the fibre discharges are synchronised to a particular frequency component. It is defined as the magnitude of the Fourier component at that frequency divided by the DC component which is the average discharge rate. Plots of synchronization index against harmonic frequency are called normalized harmonic spectra (NHS). The NHS obtained [35] varied from fibre to fibre with respect to the CF. The largest spectral component in the response patterns are harmonics that are close in frequency to a formant, the fundamental or the CF. The principal factor that determines which of these will be the largest is the relation of the fibre's CF to the formants. To portray a representation of the stimulus across frequency, fibers from regular intervals along the CF dimension can be used. For this purpose the axis of CFs was divided into bins and the response of the fibres within each bin was arranged to obtain "band-average" spectra. Note that at this stage there is a "band-average" NHS for each of the different frequency bins. These represent "typical" NHS for the range of fibres with CF falling into the corresponding bin. The next step is to map the response of the whole array of fibres tonotopically arranged. This represents the complete mapping of the stimulus onto the physiological dimension at the level of the auditory nerve. In [35] a pseudo-perspective representation was adopted where each band-average power spectrum was plotted with frequency along the oblique axis and amplitude along the vertical axis. The horizontal axis represents the different CF bins tonotopically arranged [figure 2.1-24].

At a first glance it can be seen that different patterns are obtained from each vowel. A schematic pattern is also shown in figure [2.1-25]. High responses obtain for frequencies at the CF (f = CF) and also for the formant frequencies as well as for the fundamental. Since the frequency selective elements especially at high frequencies do not resolve individual harmonics of the fundamental frequency their outputs would show considerable envelope modulation at the fundamental frequency. Rectification of this modulation at later stages of processing produces a prominent fundamental frequency component. The results from [35] show large components at F0. The larger components occur at different CFs for different vowels. Another important issue from the speech bit rate compression point of view is to what extent harmonics in regions other than the formants have an effect on the NHS. This seems to vary with how closely the formants are in the frequency domain; for formants close together little synchronisation can be seen for in between regions, but for spread vowels large components were found form fibers with CFs in the interformant regions.

Synchronisation spectra can be obtained with noise like stimuli as well. In [37] experiments were performed to describe how the spectra of certain voiceless fricatives are represented in the discharge patterns of auditory-nerve fibres. These sounds have an "incomplete" formant pattern as not all the resonances of the vocal tract are excited by the turbulence noise. Most of the energy of frigative sounds is located in specific bands of the frequency spectrum. This information could either be represented in the auditory nerve in the profile of average discharge rates against CF or alternatively (or in addition) the spectral information could be coded in the fine time patterns of fibre discharges.

White Gaussian noise was used to generate the sounds by passing them through one to three bandpass filters.

In order to estimate the profile of the average discharge rates against CF two sets of PST histograms were computed with the rather large bin width of 0.25ms. Two different windows were used to sum the histogram bins. The first, a trapezoidal weighting window beginning at the onset of the stimulus and having a central duration of 50ms and a total duration of 200ms was used to determine the "steady-state" rate. The "onset" rate was measured with a second window of value 1 from 0 to 10ms and then decreasing linearly to reach 0 at 50ms.

To estimate power spectra through synchronization indices a second set of PST histograms was computed with a small bin of 0.025 ms for the central 100 ms segments of the stimuli.

This bin width is suitable for measuring frequency components up to 8KHz. Power spectra were obtained by averaging the DFTs (magnitude) obtained from 16 overlapping 12.8ms segments of the histogram weighted by a Kaiser window. "Band averaging" amongst CFs was used as in [35]

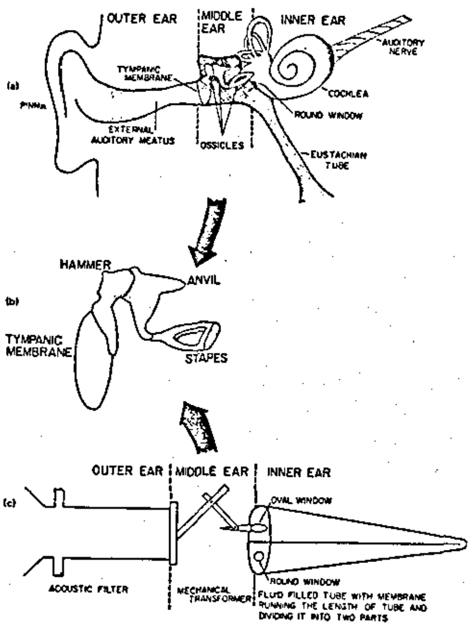
From a study of the average discharge rates the profiles against CF for either the onset or the steady state conditions provide a rough indication of the frequency regions where the stimuli have most of their energy.

Synchronization rates, through band averaged power spectra show a strong component at the CF of each fibre. Although fibre time patterns of discharge tend to differ amongst the stimuli no prominent response components are found at formant frequencies.

It seems that for the noise-like stimuli average discharge rates reflect differences in the stimuli better than data from synchronization rates. This is aided by the fact that fricatives have lower intensities than vowels, so that saturation levels of the rate-intensity functions are less likely to be reached. Also for vowels, the intense components near the first formant frequency seem to suppress (mask) the discharge rates of fibres at the higher formants. This is not so for frigatives.

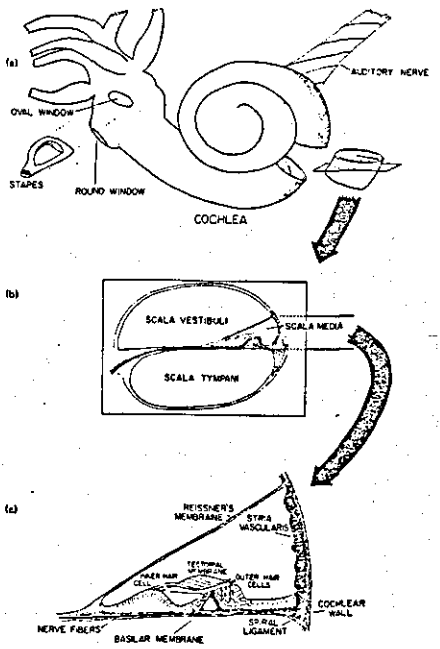
For sounds with dynamic characteristics [38], discharge rates of auditory fibres in response to formant transitions were found to be dependent on the preceding context. They also contained cues about the nature of the transitions. It is possible that peaks in discharge rate occurring in response to rapid changes in amplitude or spectrum might be used by the speech processing centre as pointers to portions of speech signals rich in phonetic information.

Cues for distinction among vowel stimuli must persist in the presence of background noise. These should reflect the ability of subjects to distinguish amongst nonsense syllables in noise. Intelligibility is high for nonsense syllables as long as the signal-to- noise ratio is above 5-10dB for broad band noise. In [39] responses of auditory-nerve fibres to steady-state, two formant vowels in low-pass background noise (SNR = 10dB) were obtained. These results showed that strong effects of noise on the fine time patterns of discharge were limited to CF regions far from formant frequencies: The discharge patterns contained many cues for distinctions among the stimuli, consistent with psychophysical performance at moderate SNRs.



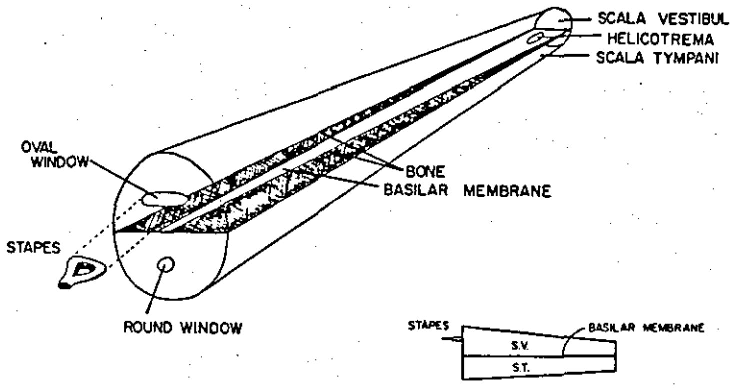
Major divisions of the peripheral auditory system: (a) the outer, middle and inner ear; (b) an enlarged version of the middle ear; (c) the mechanical analogs of the outer and middle ear.

Figure 2.1–1 [a]



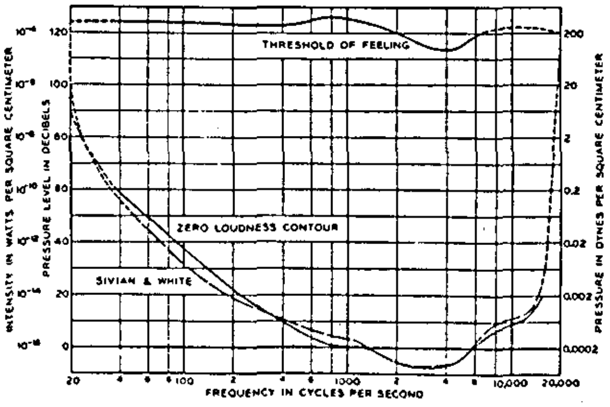
Structural and anatomical features of the cochlea: (a) the cochlea in relation to the middle ear, the vestibular channels (partially illustrated) and the auditory nerve; (b) cross section of the cochlea; (c) the structures within the scala media.

Figure 2.1–2 [a]



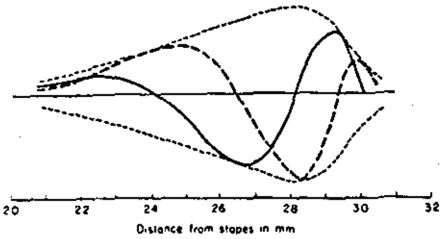
The major structural features of the uncoiled cochlea. Note that the basilar membrane is narrow near the round window and wider near the helicotrema, a taper opposite the cross-section area of the cochlea.

Figure 2.1–3 [a]



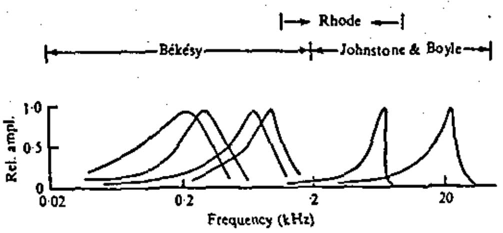
—Auditory Area Between Threshold of Feeling and the Threshold of Hearing.

Figure 2.1–4 [b]



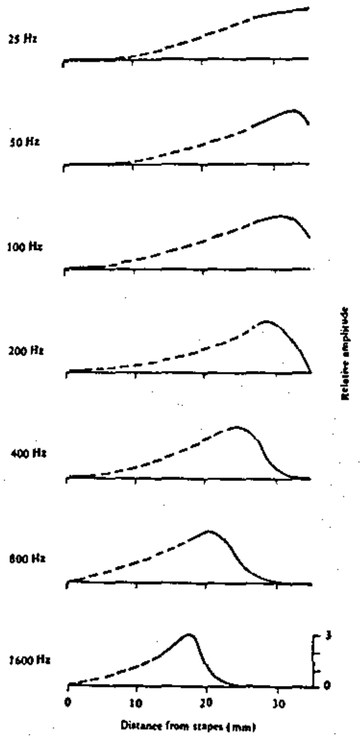
The instantaneous displacement of the cochlear partition at two successive instants in time, derived from a cochlear model. The pattern moves from left to right, building up gradually with distance, and decaying rapidly beyond the point of maximal displacement. The dotted line represents the envelope traced out by the amplitude peaks in the waveform.

Figure 2.1–5 [1]



A comparison of the shapes of 'tuning curves' on the basilar membrane obtained by three different sets of workers. Each curve was obtained by measuring the amplitude of vibration at a particular point on the basilar membrane as a function of the frequency of stimulation.

Figure 2.1–7 [d]



Envelopes of patterns of vibration on the basilar membrane for a number of low frequency sounds. Solid lines indicate the results of actual measurements, while the dashed lines are von Békésy's extrapolations. From *Experiments in Hearing* by von Bekesy, G (1960)

Figure 2.1–6 [42]

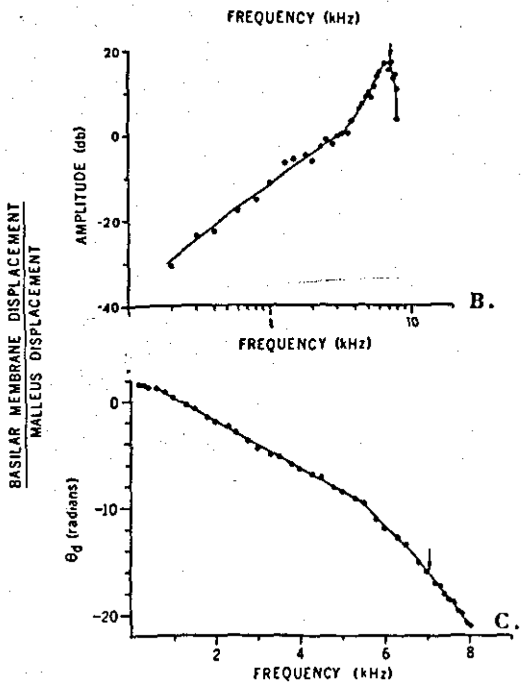
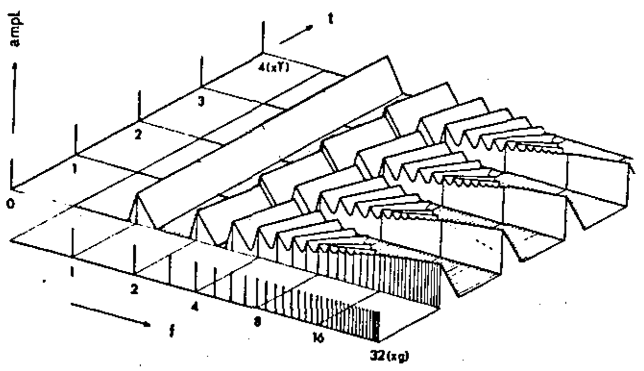
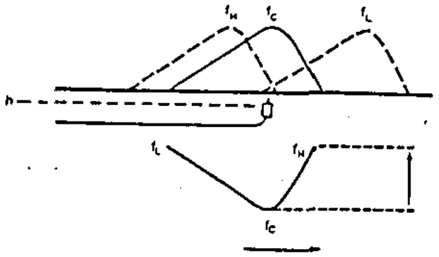


Figure 2.1–8 [4] Amplitude of vibration of the malleus and of the basilar membrane as a function of frequency. The measurements in the vicinity of the maximum ratio were made at 80 dB SPL, while those at lower frequencies were made at higher intensities and have been linearly extrapolated for a stimulus of 80 dB SPL. In this and all figures, each point is an experimentally determined value, and each line is a free-hand fit to these points. B: Input/output ratio, in decibels, for the malleus and the basilar membrane. C: Phase differences between the motion of the basilar membrane and the motion of the malleus. Negative numbers signify that the motion of the basilar membrane lags the motion of the malleus. The arrows indicate the values of the curves at the maximally effective frequency. Note that the value of the phase difference is 1.6 rad, about 90°, at frequencies less than 300 Hz (Animal 70-15).



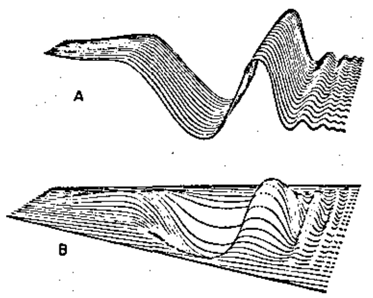
Schematic representation of the response of the basilar membrane to a series of periodic impulses. Time is plotted along the *t* axis on a linear scale; frequency (*f*) and amplitude (ampl.) scales are logarithmic. The waveform of the stimulus is indicated on the left and its spectrum is at the bottom. For the central part of the figure the frequency axis may be considered as equivalent to a position axis, indicating distance along the basilar membrane. This part of the figure shows the envelopes of the travelling wave patterns as a function of time and position

Figure 2.1-9 [d]



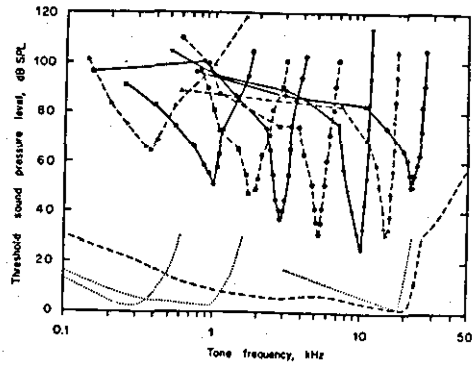
Hypothetical frequency response curve of a single hair cell (h) (radial innervation). The upper drawing depicts (full line) the vibration envelope on the basilar membrane at say 40 db above threshold for the frequency f_C which most strongly excites the hair cell. At the same intensity, the highest frequency which will just excite the hair cell is f_H, and the lowest frequency is f_L (interrupted curves). The lower drawing depicts the resultant threshold/frequency response curve for this hair cell. The threshold for frequencies f_L and f_H will, by definition, be +40 db, while the threshold at f_C will be 40 db below this level. Thus the threshold/frequency response curve mirrors the basilar membrane disturbance

Figure 2.1-10 [e]



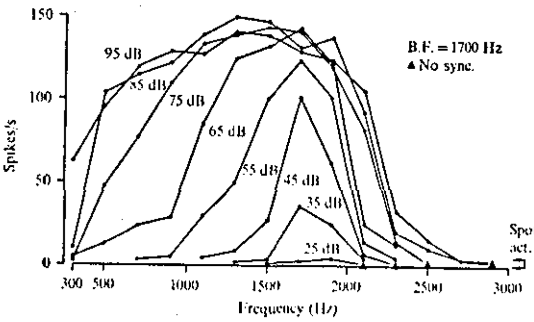
Traveling-wave patterns: (*a*) executed by a hypothetical ribbon-like partition; (*b*) observed along the single-layer partition of a cochlear model. Scales are arbitrary in both drawings, and magnitudes are exaggerated.

Figure 2.1 -11 [11]



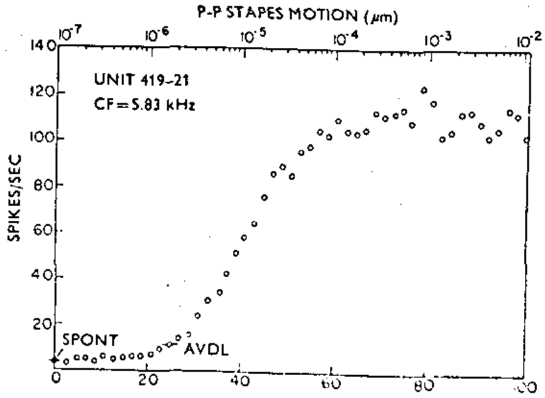
A sample of tuning curves (frequency threshold curves) of single fibres in the auditory nerve of anaesthetized cats. For each fibre the threshold is plotted as a function of stimulating frequency (logarithmic scale). The dotted and dashed curves at the bottom show corresponding measurements from the basilar membrane. The sound level required to produce a constant amplitude of vibration at a particular point on the basilar membrane is plotted as a function of frequency. The position of these curves on the ordinate is arbitrary; they have been shifted downwards for clarity.

Figure 2.1- 12 [27]



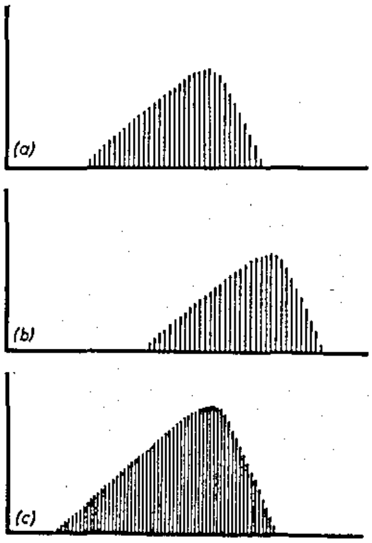
Iso-intensity contours for a single fibre in the auditory nerve of an anaesthetized squirrel monkey. Note that the frequency producing maximal firing varies as a function of level.

Figure 2.1-13 [16]



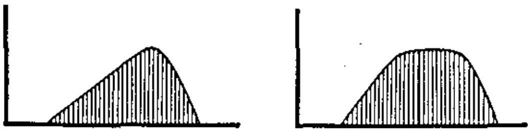
An example of how the discharge rate of a single auditory nerve fibre varies as a function of stimulus level, for a continuous stimulating tone at the CF of the neurone. The threshold of the neurone is the lowest sound level at which there is a detectable change in firing rate and is indicated by the letters AVDL (audiovisual detection level). Above a certain sound level increases in level do not produce increases in firing rate, the neurone is saturated. The range of levels between threshold and saturation is known as the 'dynamic range' and is typically 30–40 dB. The sound level is specified in dB with an arbitrary reference level. The level at the point marked AVDL corresponds to about 2 dB SPL.

Figure 2.1-14 [17]



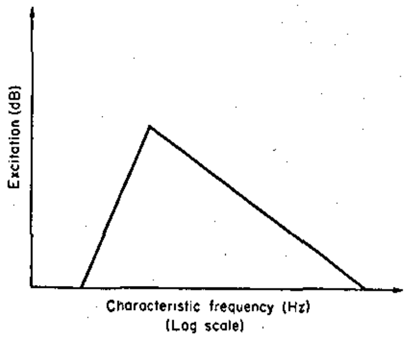
The distribution of activity in the array of auditory nerve fibres for two different stimulus frequencies (*a*), (*b*). Each vertical bar represents a small group of adjacent fibres, and its height the mean discharge rate of those fibres. (*c*) represents the response to a stimulus of the same frequency as (*a*) but of higher intensity. The response involves both higher discharge rates and more fibres.

Figure 2.1-15 [e]



The distribution of pulse rates in the active array may vary according to the intensity/rate curves of the fibres involved.

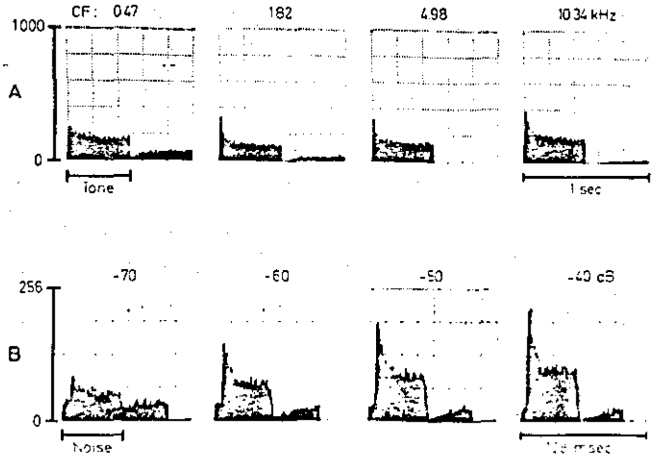
Figure 2.1-16 [e]



A schematic and idealized representation of the 'excitation pattern' evoked by a pure tone. The pattern represents the effective level of the stimulus in dB at each characteristic frequency.

Figure 2.1-17 [d]

24 E. F. EVANS: Cochlear Nerve and Cochlear Nucleus



Time-course of response of cat cochlear fibres to tone and noise bursts. PST histograms. A: 4 different fibres of the CF indicated. Linear ordinate scale, number of spikes per bin; 2 min data; tone at CF, 0.5 sec duration, presented 1 sec. B: Effect of level of stimulus on time-course (similar for tones and noise bursts). 50 msec noise burst beginning approx. 2.5 msec after zero time of PST, repetition rate: 10/sec. Relative signal level indicated above each PST histogram.

Figure 2.1-18 [27]

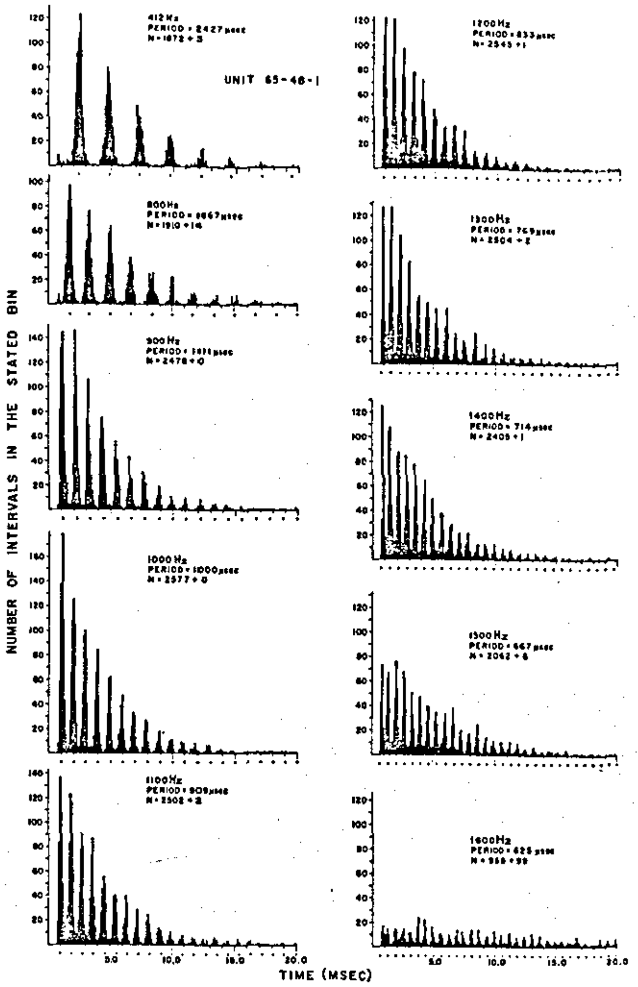
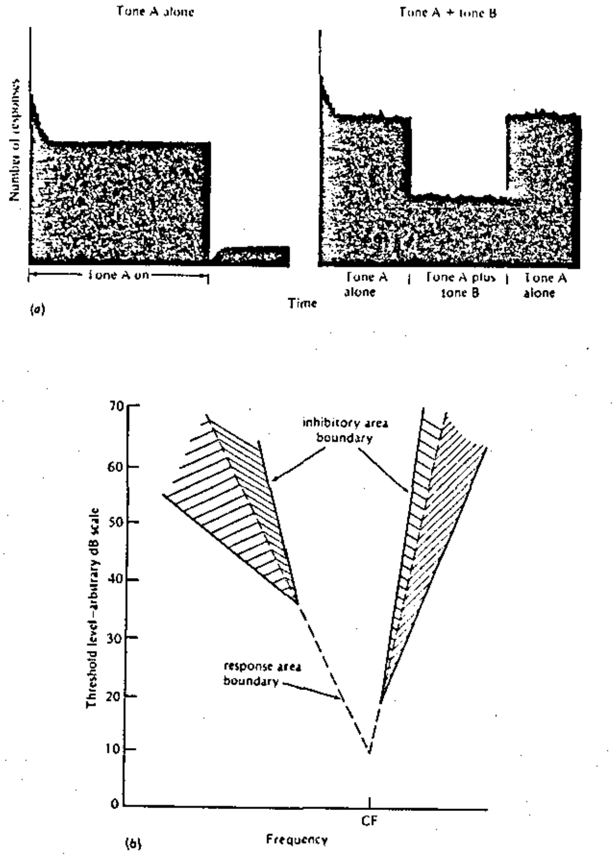
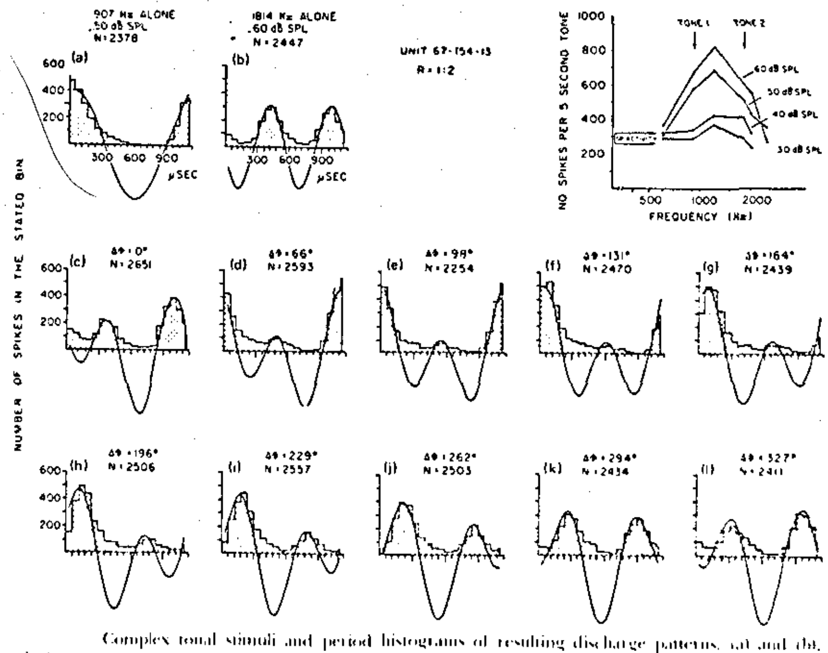


Figure 2.1–19 [28] Interspike intervals for a single auditory neuron of a squirrel monkey in response to 1 sec tones at 80 dB SPL. Stimulus frequency is shown above each graph. The dots below the abscissa are integral multiples of the period of the stimulus. (N is the number of intervals plotted plus the number of intervals with values greater than shown on the abscissa.)



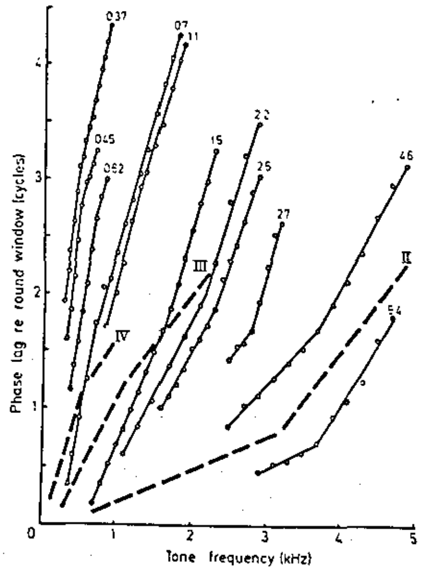
A schematic diagram showing two-tone inhibition. In the top figure (7.23a) the PST histogram to tone A presented alone is shown on the left. If a second tone of the correct frequency and intensity (tone B) is added to the first tone (tone A), the discharge rate of the nerve fiber can be reduced during the time tone B is added to tone A. The bottom figure (7.23b) depicts the typical two-tone inhibition result in terms of a tuning curve diagram. The nonhashed area shows the frequencies and intensities of tone A that excite the nerve. The hashed areas show the frequencies and intensities of tone B, which when added to tone A, decrease (inhibit) the discharge rate of the nerve responding to tone A.

Figure 2.1–20 [43]



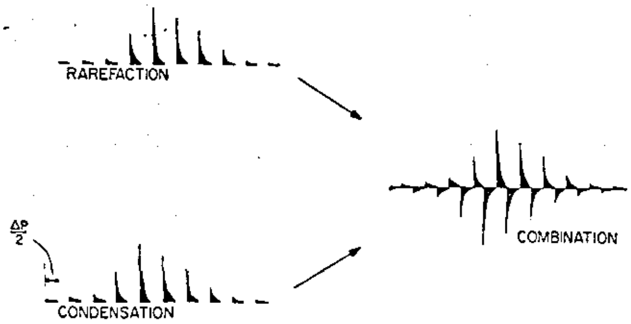
Complex tonal stimuli and period histograms of resulting discharge patterns. (a) and (b), primary tones; (c)-(l), complex tones; Δφ is phase shift between the primaries; upper right-hand graph shows the response area at various SPLs.

Figure 2.1–21 [29]



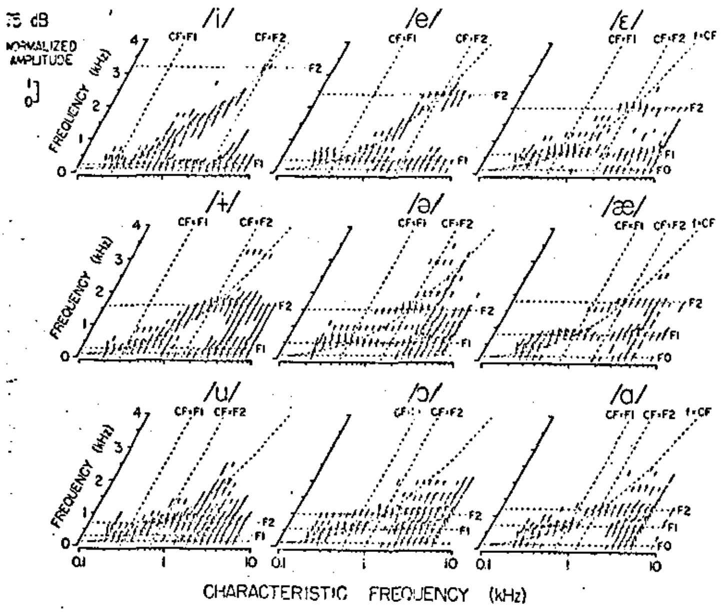
Phase of response synchronization relative to round window cochlear microphonic potential for 11 cochlear fibres from the cat. Computed by Fourier transform of period histograms obtained at different frequencies of continuous tone stimulation at constant intensity. CF of fibres (in kHz) given above each plot. Note break-point occurring at about CF for each fibre. II, III, IV: phase characteristics of cochlear microphonic potential recorded by differential electrodes in second, third, and fourth turn, respectively, of guinea pig cochlea.

Figure 2.1-22 [32]



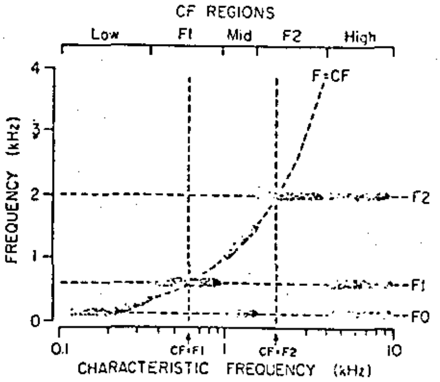
Composite of rarefaction and condensation clicks. The receptors are maximally sensitive during a particular half-cycle of stimulation, and thus, opposite polarity clicks tend to generate poststimulus histograms that are similar but delayed from one another by one-half the period of the characteristic frequency. By inverting the histogram obtained with one polarity and adding to the other histogram, a combination histogram is obtained. This construction gives us a better picture of membrane response by, in effect, removing a major nonlinearity of the transduction process.

Figure 2.1-23 [a]



Pseudo-perspective representation of normalized band-average power spectra for 0.55 oct CF bands in response to the nine vowel stimuli presented at 75 dB SPL. The normalized power spectrum is the square of the normalized harmonic spectrum. Each band-average power spectrum is plotted with frequency along the oblique axis, and amplitude along the vertical axis. Spectrum points with an amplitude lower than 0.05 are omitted for clarity. The "envelopes" of the vertical lines have been drawn for each power spectrum in order to improve visual continuity. The center frequencies of the CF bands are sampled every quarter octave. Horizontal dashed lines show the positions of the fundamental frequency $F0$ and the formant frequencies $F1$ and $F2$ along the frequency axis. Oblique dashed lines mark the places of the formant frequencies along the CF dimension. The curved dashed line is the locus of points for which frequency is equal to CF.

Figure 2.1-24 [35]



Schematic diagram of the five CF regions used in the description of responses to the vowels. The horizontal and vertical axes correspond to the horizontal and oblique axes of Fig. 6, respectively. Dark areas represent the largest response components. The dashed lines correspond to the same landmarks as in Fig. 6. For the purpose of exposition, the extents of the different CF regions are greatly exaggerated, and do not correspond to those observed for any vowel.

Figure 2.1-25 [35]

1. G.V. Békésy (1947) "The Variation of Phase Along the Basilar Membrane with Sinosoidal Vibrations", JASA vol. 19, No. 3 pp452-460.

2. D.J. Anderson, J.E. Rose, J.E. Hind, J.F. Brugge "Temporal Position of Discharges in Single Auditory Nerve Fibres within the Cycle of a Sine-Wave Stimulus: Frequency and Intensity Effects" JASA, Vol. 49, No. 4 (Part 2), 1971 pp1131-1139.

3. P. Dallos, M.A. Cheatham, "Travel Time in the Cochlea and its Determination from Cochlear-Microphonic Data", JASA, Vol. 49, No. 4 (part 2) 1971, pp1140-1143.

4. W.S. Rhode, "Observations of the Vibration of the Basilar Membrane in Squirrel Monkeys using the Mössbauer Technique", JASA, Vol. 49, No. 4 (Part 2) 1971 pp1218-1231.

5. W.S. Rhode, "An Investigation of post-Mortem Cochlear Mechanics using the Mössbauer Effect". In "Basic Mechanisms in Hearing", pp49-67. A.R. Moller Ed. New Work: Academic Press, 1973.

6. M.R. Schroeder "Models of Hearing" IEEE proc. Vol. 63, No. 9, Sept. 1975 pp1332-1350.

7. J.L. Flanagan "Speech Analysis Synthesis and Perception" Springer-Verlag-Berlin - Heidelberg, New York 1972.

8. W.S. Rhode "Vibration of the Basilar Membrane observed with the Mössbauer technique" in "Physiology of the Auditory System" B. Sachs Ed. pp31-37, 1971.National Educational Consultants Inc

9. J. Tonndorf: Discussion on cochlear Mechanisms. In Physiology of the Auditory System B. Sachs Ed. pp73, 1971, National Educational Consultants Inc.

10. H. Duifhuis "Cochlear nonlinearity and second filter: Possible mechanism and implications" J. Acoust. Soc. Am. Vol. 59, No. 2, Feb. 1976. pp408-423

11. J. Tonndorf "Cochlear Mechanics and Hydrodynamics" in Foundation of Modern Auditory Theory" J.V. Tobias ed. Academic Press, 1970. pp203-250

12. J. Tonndorf "The Significance of shearing Displacements for the Mechanical Stimulation of Cochlear hair cells" in Facts and Models in Hearing E. Zwicker and E. Terhardt eds. Springer-Verlag, 1974 pp65-75.

13. J.J. Zwiszocki and W.G. Sokolich "Neuro-mechanical frequency analysis in the cochlea" in Facts and Models in Hearing, E. Zwicker and E. Terhardt eds. Springer-Verlag. 1974, pp107-117

14. H. Duifhuis "Cochlear nonlinearity and second filter: Possible mechanisms and implications" J. Acoust. Soc. Am. Vol. 59, No. 2, Feb. 1976, pp408-423.

15. Kiang, N.Y.-S, Watanabe, T, Thomas, E.C. Clark, L.F., "Discharge Patterns of Single Fibers in the Cat's Auditory Nerve", MIT Press, Cambridge, Massachusetts. (1965). No. 35 (Research Monograph)

16. J.E. Rose, J.E. Hind, D.J. Anderson, J.F. Brugge, "Some effects of stimulus intensity on the response of auditory nerve fibers in the squirrel monkey", J. Neurophysiol 34, 685-699, (1971).

17. Kiang, N.Y.-S, "A Survey of recent developments in the study of Auditory Psychology". Ann. Otol. Rhinol. Laryngol. 77, pp656-675 (1968).

18. M.C. Liberman, "Auditory nerve response from Cats raised in a low-noise chamber, J. Acoust. Soc. Am. 63 pp442-455, (1978).

19. M.B. Sachs, Abbas, P.J. "Rate versus level functions for auditory-nerve fibers in cats: tone-burst stimuli, J. Acoust, Soc. Am. 56, pp1835-1847. (1974).

20. J.E. Hind, J.E. Rose, J.F. Brugge, D.J. Anderson, "Some effects of intensity on the discharges of auditory nerve fibers" in 'Physiology of the Auditory System', M.B. Sachs ed. National Educational Consultants (1971). pp101-111.

21. J.P. Wilson "Basilar membrane vibration data and their relation to theories of frequency analysis" in Facts and Models in Hearing, E. Zwicker and E. Terhardt eds. Springer-Verlag (1974), pp56-63.

22. E.F. Evans, J.P. Wilson, "Frequency sharpening of the Cochlea: The effective bandwidth of Cochlear nerve fibres" in Seventh International Congress on acoustics, 23H4, pp453-456, Budapest (1971).

23. E.F. Evans "Peripheral Processing of Complex sounds" in Dahlem Workshop of Complex Acoustic Signals pp146-159 (1977)

24. Evans, E.F. 1975 "Cochlear nerve and Cochlear nucleus" In Handbook of Sensory Physiology, Vol. V, Part II, ed. W.D. Keidel and W.D. Neff, pp1-108, Berlin Heidelberg New York: Springer-Verlag.

25. R.R. Pfeiffer and D.O. Kim (1975) Cochlear nerve fiber responses: Distribution along the cochlear partition, J. Acoust. Soc. Am. 58, 867-869.

26. Sachs, M.B. and Young, E.D. (1980) Effects of nonlinearities on speech encoding in the auditory nerve, J. Acoust. Soc. Am. 68, 858-875.

27. E.F. Evans "Cochlear Nerve and Cochlear Nucleus" in Handbook of Sensory Physiology" Volume V/2 W.D. Keidel and W.D. Neff eds. Springer-Verlag Berlin, Heidelberg, New York, 1975. pp1-108

28. J.E. Rose, J.F. Brugge, D.J. Anderson, J.E. Hind "Phase locked response to low frequency tones in the single auditory fibres of the squirrel monkey", Journal of Neurophysiology, 1967, 30, 769-793.

29. J.F. Brugge, D.J. Anderson, D.J. Hind, J.E. Rose, "Time structure of discharges in single auditory nerve fibers of the squirrel monkey in response to complex periodic sounds, Journal of Neurophysiology, 1969, 32, pp386-401.

30. Rose, J.E. "Discharges of single fibers in the Mammalian Auditory Nerve" in Frequency Analysis and Periodicity Detection in Hearing" R. Plomp and G.F. Smoorenburg eds. A.W.Sijthoff * Leiden * 1970. pp176-192

31. Hind, J.E., Anderson, D.J., Brugge, J.F., Rose, J.G. "Coding of information pertaining to paired low-fequency tones in sinle auditory nerve fibres of the squirrel monkey, J. Neurophysiology 30, pp794-816, 1967.

32. Pfeiffer, R.R. and Molnar, G.E. "Cochlear nerve fibre discharge patterns: relationship to the cochlear microphonic", Science 167, pp1614-1616 (1970).

33. Goldstein, J.L., Baer, T., Kiang, N.Y.S. "A theoretical treatment of latency, group delay and tuning characteristics for auditory-nerve responses to clicks and tones" In: Physiology of the Auditory System, pp133-141 Baltimore, Md: National Educational Consultants, 1971.

34. Goblick, T.J. and Pfeiffer, R.R. "Time-domain measurements of the cochlear nonlinearities using combinations click stimuli", J. Acoust, Soc. Am. 46, pp924-938, 1969.

35. B.Delgutte and N.Y.S. Kiang "Speech coding in the auditory nerve. I, Vowel-like sounds". J. Acoust. Soc. Am. 75(3), March 1984, pp866-878.

36. B. Delgutte and N.Y.S. Kiang "Speech coding in the auditory nerve. II. Processing schemes for vowel-like sounds". J. Acoust. Soc. Am. 75(3) March 1984, pp879-886.

37. B. Delgutte and N.Y.S. Kiang "Speech coding in the auditory nerve III. Voiceless frigative consonants, J. Acoust. Soc. Am. 75(3) March 1984, pp887-896.

38. B. Delgutte and N.Y.S. Kiang "Speech coding in the auditory nerve IV. Sounds with consonant-like characteristics", J. Acoust Soc. Am. 75(3) March 1984, pp897-907.

39. B. Delgutte and N.Y.S. Kiang "Speech coding in the auditory nerve to Vowels in background noise", J. Acoust. Soc. Am. 75(3) March 1984 pp908-918.

40. Sachs, M.B., Young, E.D. "Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate", J. Acoust. Soc. Am. (66) 470-479, 1979.

41. Young, E.D. and Sachs, M.B. "Representation of steady state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers", J. Acoust. Soc. Am. 66 1381-1403. 1979

42. Békésy, G. Von, Experiments in hearing (trans. and ed. E.G. Wever) 1960 McGraw-Hill, New York.

43. W.A. Yost, D.W. Nielsen "Fundamentals of Hearing" Holt, Rinehart and Winston, 1985.

a) Green, D.M., "An Introduction to Hearing", Lawrence Erlbaum Associates; Hillsdale, New Jersey (1976).

b) Fletcher, H. "Speech and Hearing in Communication", D.Van Nostrand Company"; Princeton New Jersey (1958).

c) Schroeder, M.R., "In Recognition of Complex Acoustic Signals", Life Sciences Research, Report 5, edited by T.H. Bullock (Dahlem Konferenten) (Abakon Verlag, Berlin) pp323-328 (1977).

d) Moore, B.C.J. "An Introduction to the Psychology of Hearing", Academic Press, London (1982).

e) Whitfield, I.C., "The Auditory Pathway", Edward Arnold, London (1967).

f) Hall, J.L. "In Hearing Research and Theory", volume 1, edited by J.V. Tobias and Earl D. Schubert, Academic Press, London (1981) pp.1-61.

2.2 MODELS OF PHYSIOLOGICAL CHARACTERISTICS OF THE PERIPHERAL AUDITORY SYSTEM

2.2.1 MODELS OF HYDROMECHANICS:

Models of hydromechanics aim to describe the modes of vibration of the BM in response to various stimuli. Several different approaches can be identified. Some workers [1,2] approached the problem by constructing mechanical models of the cochlea. These constructions are necessarily many times larger than the cochlea itself. To make a useful model, the real structure has to be stripped to its essentials and thus be simplified, yet retain enough to have a functional model. After deciding upon the appropriate important parameters, dimensional analysis has to be applied. This allows the comparison of different-sized similar structures in their response to a given form of stimulation. The resulting "model" often bears little physical resemblance to the original structure: The functionality of the model stems from its similarity of responses with the original. Through these models the nature of the fluid motion in the cochlea could be studied and its effect on the BM vibration analysed. The advantage of mechanical models is that the response can be observed directly as opposed to mathematical models where the results are sometimes obscured in the equations. The increased use of computers in recent years has changed this situation since extensive results can be obtained in graphical form which enables direct observation. Hence the more recent models were mathematical ones [3].

Mathematical models themselves are divided into two categories. First the "physical" models use anatomical and physical data from the inner ear, in conjunction with the laws of physics (hydrodynamics) to describe the interactions of the various parts. In essence these models are no more than a mathematical description (through the laws of physics) of the cochlea itself. By contrast the other type of model (the "computational" model) only attempts to model the function of the cochlea, i.e. its input/output characteristics, with no necessary connection with actual anatomical and physical properties or laws.

Physical models tend to exhibit behaviour which is qualitatively similar to experimental observations. No model to date though has managed to describe quantitatively all the experimentally derived characteristics of the BM vibration. One example which serves to highlight the shortcomings of the models is the size of BM displacements predicted at threshold stimulation. The largest value is that obtained by Rhode [4] of 0.06Å when the molecular dimensions involved are of the order of 1Å (1Å = 10^{-10}m)!

Computational models have the advantage of being able to predict responses to stimuli other than those used to formulate the model. One such example is Flanagan's model [5]. Disagreements of that model with experimental data now seem to be due to experimental errors, since more recent experimental data [4] confirmed Flanagan's results.

A brief description of most mathematical models can be found in [3]. More recent models can be found in [6, 7, 8].

2.2.2 Models incorporating mechanical to neural transduction mechanisms

The mechanisms of mechanical to neural transduction are not known to their entirety. It is not known for example whether BM displacement or some derivative of it is the input to the neural transducers.

An important distinction is that of "place" theories as opposed to "periodicity" (or volley), theories. In the former all the information is encoded in the location of the maximal vibrations on the BM. The mean discharge rate versus characteristic frequency is the relevant variable. One problem stemming from this model is the limitation on the range of intensities that can be coded due to saturation of the fibers. For the volley theory the frequency of sinosoidal vibrations is coded not by place but by the periodicity in the neural response. The intensity is assumed to be coded in the number of fibers responding to a particular stimulus. Since not all fibers need be firing on every cycle of the stimulus the refractory period of a fiber need not be a limitation to the model. The main variable here is the degree of synchronization to the stimulus over all the fibers. Some degree of phase-locking is essential to this model so there is some controversy as to how frequencies above about 5KHz are coded since for these high frequencies phase locking is not observed in the nerve responses.

Specific models of physiological behaviour tend to draw attention to either mean discharge rate or synchronization index according to which model the particular author(s) support. It should be noted though that it is not necessary for a model to rely on either place or volley theory alone. It is quite possible that both theories play some role in the coding of auditory signals.

In [9] the linear part of the model comprises of two filters. One is due to the BM displacement and is generally of a low pass nature and the other is a narrow bandpass "second" filter. Within 40 dB of the threshold the response of the fibers can be described as linear.

Therefore responses can be predicted from the effective bandwidths of the tuning curves and the shape (especially near the tip) of the fiber's tuning curves. A complex signal is then analysed by the equivalent of a filterbank. Frequency components of the stimulus are encoded in the discharge rates (place theory) and for low frequencies (<5KHz) in the temporal patterns of the discharge (periodicity theory). For low frequency signals (<5KHz) the waveform and for low and high frequencies the envelope of modulation are represented in the temporal patterns through phase locking. As we have seen earlier in "phase-locking" the probability of discharge is approximately linearly related to the filtered half-wave rectified signal. The time "jitter" present in this signal limits the temporal resolution to corresponding frequencies of 4-5KHz.

A physiological model must also incorporate the various nonlinearities in the transduction process: Saturation limits to less than 50 dB the range of intensities that can be coded in terms of mean square rate. (normal speech levels are around 70 dB above threshold). As this is a serious problem of place coding of multicomponent stimuli information may be coded in the fine structure.

Another nonlinearity is that arising in two tone stimulation. The suppression effects arising in these two tone masking experiments seems to have a small effect on multicomponent signals.

The stages in a model to predict linear and nonlinear behaviour could then be as follows [9]:

1. A linear bandpass filter (to simulate cochlear filtering characteristics represented in the fiber's tuning curves).

2. A half-wave rectifier (to simulate excitation by unidirectional excursions of the BM).

3. Logarithmic amplifier (to portray a linear relation between mean discharge rate and the logarithm of signal voltage above threshold).

4. An amplitude limiter (to produce saturations of mean discharge rate).

5. AC coupling (to simulate adaptation and off suppression)

6. Probabilistic transform (to produce a linear relationship between probability of discharge and signal amplitude with spontaneous discharge).

7. A monostable with dead time (to simulate action potential spikes and their refractory period).

In addition stage 1 could be split into a lowpass filter and a bandpass filter with a cubic nonlinearity "sandwiched" between the two to produce two-tone suppression results.

Modelling of higher stages of the auditory system, higher than the level of auditory fibers would involve more specialized processing such as selectivity for direction and rate of change of frequency for the detection of frequency modulation.

More work on models can be found in [7, 8, 10, 11, 12, 13, 14], whereas specific applications of such models have been implemented in [15, 16, 17, 18, 19, 20].

Also interesting is the work presented in [21] which uses linear filtering theory to interrelate latency, group delay and tuning characteristics of auditory fibers.

1. G.V. Békézy "Enlarged Mechanical Model of the Cochlea with nerve supply". in: Tobias, J.V. (Ed.) Foundations of modern auditory theory. New York.: Academic Press, 1970. pp305-340

2. Tonndorf, J.: Cochlear mechanics and hydro-dynamics. In: Tobias, J.V. (Ed.): Foundations of modern auditory theory, New York: Academic Press, 1970. pp203-250

3. Geisler, C.D. "Mathematical Models of the Mechanics of the inner ear" in: "Handbook of Sensory Physiology", W.D. Keidel and W.D. Neff (Eds.), Springer-Verlag, 1976. V/3, pp391-415

4. W.S. Rhode "Observations of the Vibration of the Bassilar Membrane in Squirrel Monkeys using the Mössbauer Technique". J. Acoust. Soc. Am. Vol. 49, No 4(Part 2) 1971. pp1218-1231

5. J.L.. Flanagan "Speech Analysis Sythesis and Perception", Springer-Verlag, 1972 pp108-140.

6. M.R. Schroeder "An integrable model for the basilar membrane" In: J. Acoust. Soc. Am. Vol. 53, No. 2, 1973 pp429-434.

7. M.R. Schroeder "Models of Hearing" IEEE proc. Vol. 63, No, 9, Sept. 1975 pp1332-1349.

8. J.L. Hall, "Observations on a nonlinear Model for Motion of the Basilar membrane" in: Hearing Research Academic Press, 1981, pp1-61.

9. E.F Evans "Peripheral Processing of Complex Sounds" in Dahlem Workshop on Recognition of Complex acoustic Signals 1977? pp145-159.

10. H. Duifhuis "Consequences of peripheral frequency selectivity for nonsimultaneous masking" J. Acoust. Soc. Am. Vol. 54, No. 6, 1973, pp1471-1488.

11. H. Duifhuis "Cochlear nonlinearity and second filter: Possible mechanisms and implications" J. Acoust. Soc. Am., Vol. 59, No. 2, Feb. 1976 pp408-423.

12. M.R. Schroeder, J.L. Hall "Model for mechanical to neural transduction in the auditory receptor", J. Acoust. Soc. Am., Vol. 55, May 1974, pp1055-1060.

13. E. de Boer, H.R. de Jongh "Computer simulation of cochlear filtering" in seventh international congress on acoustics; Budapest 1971 20H12 pp393-396.

14. J.L. Zwislocki and W.G. Sokolich, "Neuromechanical frequency analysis in the cochlea" In: "Facts and models in Hearing" (E. Zwicker and E. Terhardt eds.) pp107-117, Springer-Verlag.

15. Y. Dologlou and J.M. Dolmazon, "Comparison of a model of the peripheral auditory System and L.P.C. analysis in a Speech Recognition System" ICASSP 1984 17.10.1-17.10.4.

16. D. Van Compernolle "A Computational Model of the Cochlea used with Cochlear Prosthesis Patients", ICASSP 1985 11.14.1-11.14.3 pp427-429.

17. O. Ghitza "A measure of In-Synchrony Regions in the Auditory Nerve firing Paterns as a basis for speech Vocoding. ICASSP 1985 13.9.1-13.9.4 pp505-508.

18. S. Steneff "Pitch and spectral estimation of Speech based on auditory synchrony model" ICASSP 1984 36.2.1-36.2.4.

19. B. Delgutte "Speech coding in the auditory nerve II Processing Schemes for Vowel like sounds". J. Acoust. Soc.Am. 75(3) March, 1984 pp879-886.

20. M.J. Hunt and C. Lefebure "Speech Recognition using a Cochlear model" ICASSP 86, TOKYO, pp1979-1982.

21. Goldstein, J.L., Baer, T., Kiang, N.Y.-S. "A theoretical treatment of latency, group delay and tuning characteristics for auditory-nerve responses to clicks and tones" In: Physiology of the Auditory System pp133-141, Baltimore, Md: National Educational Consultants, 1971.

2.3 PSYCHOPHYSICAL OBSERVATIONS AND MEASUREMENTS

2.3.1 Loudness of simple sounds (tones and narrow bands of noise)

2.3.1.1 Loudness level

The loudness level of any sound is the intensity level of the equally loud (subjectively determined) reference tone (1KHz). The unit of loudness level is the phon [d]. e.g. if a sound is subjectively assessed to be equally loud to a 1000 Hz tone at 65dB SPL then this sound has a loudness of 65 phons. In this fashion the loudness level of pure tones at various fixed intensity levels over all frequencies can be determined. What is normally done is to determine the intensity level of tones at fixed loudness levels over all frequencies or alternatively, the sensation level of tones at fixed loudness levels over all frequencies. Two such plots are given in figures 2.3-1 and 2.3-2. Since the curves in the above figures are essentially flat, for the range between 500 and 5000 Hz, then for this range the intensity level and loudness level are equal.

2.3.1.2 Loudness

The loudness level helps in constructing a scale for loudness: One starts with a standard loudness (1000 Hz tone at 40 dB SPL). This represents 1 unit of loudness, (i.e. 1 unit of loudness = 40 phons). The observer is asked to increase the level of the tone until it is judged to be twice as loud. This represents 2 units of loudness. Other values of loudness can be obtained by similar procedures. The unit of loudness is the sone. It should be noted that the loudness of a sound heard binaually is judged to be twice as loud as the same sound presented monaurally.

Various workers have attempted to produce a relationship between loudness and loudness level of simple sounds. This has been established to be a power law. Stevens [3, a] suggested the following formula

$$L = 10^{0.03P-1.2} \qquad 2.3\text{-}1$$

where L is the loudness in sones and P the loudness level in phons. This gives L = 1 for P = 40 dB. A simple rule is that the (perceived) loudness doubles when the intensity is increased by 10 dB. In terms of intensity (I) ($P^{\sim}$ 10log(I)), from above

$$L = (I/Io)^{0.3} \qquad 2.3\text{-}2$$

Io is the reference intensity corresponding to 40 dB SPL. This value of loudness refers to the sound heard binaurally.

2.3.1.3 Masking effects

It is a common experience that when any sound is impressed upon the ear it reduces the ability of the ear to sense other sounds. If while a sound A is being impressed upon the ear, another sound B is gradually increased until the sound A can no longer be heard, the sound A (maskee) is said to be masked by sound B (masker).

Masking is usually presented as a (masked) audiogram which shows the threshold of the maskee over the frequency range considered in the presence of the fixed masker, in the same way threshold in the absence of a masker is presented. It is believed in fact that this absolute threshold is due to internal masking (inside the ear), the masker being noise of cardiovascular origin (On the low frequency side).

Any sound of intensity below the threshold has a loudness of zero by definition. Such an audiogram is shown in figure 2.3-3. The lowest curve in the absolute threshold. Of the upper curves the continuous one in the masked threshold for a pure tone masker, the dashed one is for a noise band, and the dash double dot gives the frequency characteristics of the noise band. For simplicity the masked audiograms are usually plotted for the threshold shift (masked minus absolute threshold) or masking, figure 2.3-4.

2.3.1.4 Shape of the loudness curve near threshold

The function relating subjective loudness to sound pressure was shown to be a power function:

$$N = K(I)^{b} \qquad 2.3\text{-}3$$

where N is the magnitude of the sensation and I the stimulus magnitude.

Although a power function relation can be obtained from most observers in general, there is a variability concerning the exponent of the function and the values quoted in a previous section represent a statistical average. The average exponent from 10 or 12 observers seems to be quite stable and reproducible [1].

The loudness function departs from a power law near threshold, whether this is the absolute threshold or the masked threshold. The departure from a power law starts at a sensation level of about 40 dB. Through very elaborate experiments to reduce variability amongst the data [2] it was found that the loudness curve approximated a power function between loudness and sensation level (dB above threshold) between 30 and 100 dB SL. Below 30 dB it becomes progressively steeper and at around 4 dB it approached proportionality between loudness and sound intensity. The exponent of the power low was found to be 0.54 between loudness and sound pressure. Note that whether the sound is presented monaurally or binaurally does not affect the exponent but only the constant of proportionality between the power function and the loudness.

In [2] the shape of the curve for loudness near threshold was determined but no attempt was made to provide an analytical formula.

In [3] the argument is followed that if the relationship $N = K(I)^b$ represents the fundamental relation between stimulus and loudness, the reduction in loudness at threshold is equal to $N_o = K(Io)^b$ where Io is the intensity of the sound at its threshold (i.e. when N = 0). This is so, since its loudness at intensity Io should have been $N_o = K(Io)^b$. Since the loudness is actually zero, the reduction is given by the above formula. This reduction was attributed to masking from physiological noise in the ear itself. Further, the assumption was made that this reduction is constant throughout the sound's audible range, hence the resultant loudness should be given by

$$N = K(I)^b - K(Io)^b \qquad 2.3\text{-}4a$$

$$= K\ (I^b - Io^b) \quad (I > Io\text{, zero otherwise}) \qquad 2.3\text{-}4b$$

This equation was found to fit the results of [2] remarkably well [figure 2.3-5].

When, in addition to the internal physiological noise, external noise is present, there is additional masking due to the external noise. The original assumption that the internal noise is responsible for the reduction in loudness and the fact that the internal and external noises are uncorrelated allows one to equate Io in the above equation to (Ip + Ie) where Ip is the intensity of the internal noise and Ie the intensity of the external noise. Hence the general form of the loudness function can be given by:

$$N = K [I^b - (Ip + Ie)^b] \qquad 2.3\text{-}5$$

experimental results presented in [3] figure [2.3-6] showed a general agreement with the above equation. Further experiments reported in [4] though failed to show agreement with the above equation. The difference between binaural (used in [3]) and monaural [in 4] masking may account for the discrepancy.

2.3.1.5 <u>Masking of Pure tones by Pure Tones</u>

To obtain the effects of masking of one tone (the maskee) by another tone (the masker), the masker tone is kept at a constant sensation level while the maskee tone is gradually increased in intensity until it is just perceptible in the presence of the masker tone. The level, expressed in decibels, that the maskee tone is raised above its threshold level in the quiet is called the threshold shift or the masking. The results of such measurements are shown in figure [2.3-7]. The frequency of vibration of the masker is given by the number at the top of each chart and its sensation level by the number on each curve. The frequency of the maskee is given by the abscissa and its threshold shift (masking) by the ordinate.

The most prominant characteristic of the curves is the asymmetry in masking, especially at high masker levels: the masking is a lot more effective for low frequency maskers and high frequency maskees than vice versa. The asymmetry is increased as the intensity of the masker is increased.

The peculiar shape of the curve around the frequencies where the maskee frequency is close to the masker frequency or its harmonics is primarily due to beats. Beats occur at the harmonics due to nonlinearities of the auditory periphery, figure [2.3-8].

The results of figure [2.3-7] can be plotted in a different way [figure 2.3-9]. The abscissae represent the sensation level of the masker tones whose frequency is indicated near the top of each of the charts. The amounts of masking are plotted as ordinates. The most important effect that can be seen from these figures is that the curves for different frequencies intersect. This implies that if a complex sound is changed in intensity then the prominance of its different components will also change: The sensation produced by a complex sound is different in character when the sound is increased in intensity. In general as the tone becomes more intense, the low tones will become more prominent because the high tones are masked. Although this effect will be different when all the tones are sounded simultaneously, as the data were taken for two tones only, the general picture will still be true.

The implications for speech coding are as follows: Coding noise at lower frequencies will become more prominent as opposed to coding noise at high frequencies, when the intensity of the decoded speech is increased in any subjective test.

2.3.2 Physiological correlates - The critical band

When either the masker or the maskee have a broader spectrum than that used in the above experiments the effects become more complicated. The shape of the stimuli spectra come into play through the selectivity of the ear itself. Even from the masking curves for pure toes, the effects of the selectivity of the system can be seen.

These are closely related to the pattern of aural activity at the cochlea [5, 6]. Recently, psychoacoustical tuning curves have been obtained [7]. To obtain a physiological tuning curve the SPL of a tone is increased until the spike rate of a single fibre reaches a particular value just above spontaneous activity. Since a single fiber is strongly related and is maximally sensitive to a particular characteristic frequency (CF) it can be simulated psychoacoustically by using a very faint pure tone (called a CF-tone) at the characteristic frequency. The role of the tone used to obtain the physiological tuning curves is taken by another tone (called the f-tone) which (just) masks the CF tone. The SPL of the f-tone necessary to just mask the CF tone as a function of the frequency f represents the psychoacoustical tuning curve. The analogy with the physiological tuning curve comes about through the following argument: Since the CF-tone is very faint it produces little (amount of) excitation which is confined around a small area on the BM and thus excites a small number of fibers. Hence one can consider the faint CF-tone to excite only one hypothetical fiber with CF at the frequency of the CF-tone, to a first approximation. Since the f-tone just masks the CF-tone its excitation at the CF of the hypothetical fiber is also small. The SPL at which the f-tone just masks the (very faint) CF-tone can therefore be considered as that SPL which just produces a response on the hypothetical CF fiber. Hence the variaton of this SPL with frequency is directly analogous to the variation of SPL in the case of physiological tuning curves. The similarity between the physiological and psychophysical curves obtained is obvious [figure 2.3-10, 2.3-11]. This similarity was also obtained in [8].

When the masker has a broader spectrum than a tone, the effects of beats are not so pronounced and measurements can be made more accurately. Experiments with narrow noise bands [9] also seemed to confirm that masking, can be related to the spread of cochlear activity as the intensity of the masker is increased above the threshold.

The concept of an auditory filter becomes more relevant as the bandwidth of the masker increases: If the effect of the maskee is confined to a certain region on the BM and hence to a range of fibers with CFs near the stimulus then the effect of the masker is significant only over this region of CFs. If the bandwidth of the masker is increased beyond this effective region, the masking effect should not change since it does not affect the excitation due to the maskee any more than it did before. In other words, only the components of the noise falling within the bandwidth of the auditory filter will have an effect upon the detection of the maskee tone. It is assumed here that the auditory filter is centered on the stimulus one is trying to detect. Indications in favour to the above argument were found in masking experiments with wideband (white) noise [5]. Fletcher found that only a limited part of a wide band (white) noise was actually instrumental in masking a tone presented in noise. Fletcher made the assumption that this band of noise was just wide enough to contain an amount of power equal to the power of the just masked tone. On this assumption a "critical" bandwidth could be calculated from the ratio of the power of the signal at masked threshold to the power of 1 Hz-wide band of noise. This was later found to be about 40% as wide as the bandwidth of the auditory filter and was later termed the "critical ratio" reserving the name "critical bandwidth" for the actual width of the auditory filter [10]. To measure the size of the critical bandwidth and to provide further proof for the concept of an auditory filter, Greenwood [11] measured masked thresholds of pure tone signals in the presence of bands of masking noise. Detailed audiograms were mapped out by measuring the threshold of several frequencies in and near the band of noise. Following the assumptions of an auditory filter, the masked threshold of a tone at the center of the band will not increase as the bandwidth grows beyond critical width. Furthermore, since an extension of the band beyond the critical width will add "surplus" noise components on each side of the critical band, there will exist within the limits of a supracritical band of noise a frequency interval over which the signal can be varied without changing its masked threshold. This follows because, within a supracritical band, a tone need not be at the center for it to be surrounded by a critical band of noise. In other words audiograms produced by critical and supracritical bands of noise should produce respectively, "triangular" and "trapezoidal" masked audiograms. This technique enables the calculation of critical bandwidth without the need to specify the center of the critical band. Typical results are shown in [figure 2.3-12]. A narrow band masking stimulus produces a triangular audiogram. As its bandwidth, and therefore its total power is increased, the height of the audiogram increases by an amount that approximates the increase in total power. As long as the noise bandwidth is less than critical, the height continues to be proportional to the total power in the masking noise. When the bandwidth of the masking noise exceeds a critical bandwidth, the audiogram broadens into a trapezoidal form without further increase in height. The values of critical bandwidths were found to be about the same as those measured in other experiments [12].

When the audiograms are plotted on the critical band scale, the width of the flat top of a trapezoidal audiogram is approximately equal to the amount, by which the supracritical band exceeds critical bandwidth. In general, masking stimuli that are equal fractions of multiples of a critical band produce comparable audiograms anywhere in the frequency range investigated i.e. from 400-4500 Hz. The results showed that in any part of the frequency scale, masked threshold is proportional to the total power within a critical band. When the amount of masking is plotted as a function of the level of the masking stimulus, functions with slope equal to unity are obtained [figure 2.3-13]. The vertical jog shown in the figures occurs when a transition level is reached. At this level the triangular audiograms abruptly flatten. This seems to happen for both subcritical and subracritical noise bands. Nevertheless, the critical bandwidth appears to be independent of the level of the masking stimulus.

Greenwood also used pure tones to mask narrow (60 Hz) bands of noise. These experiments are more relevant to the field of speech coding where the speech harmonics would mask coding noise. The audiograms of one and two tone maskers were similar to the audiograms of subcritical noise maskers of equal power apart from the fact that they were about 3 dB lower. The three tone audiogram was nearly identical in height as well as shape to the audiogram produced by noise [figure 2.3-14].

The effects of moving the two tones in the two tone masker case can be seen in [figure 2.3-15]. The form is triangular when the tones are close together, rounded as the tones are moved further apart and double peaked when the tones exceed critical separation. For these experiments fairly low level tones were used (~ 50 dB). At high levels [figure 2.3-16] some irregularities appear in the audiogram. Also the peak of the audiogram is about 10 dB lower than that produced by a critical band of noise of equal power. This asymmetry of masking between noise and tone will be addressed again later. To summarize Greenwood's results, the data support the hypothesis that a tone at threshold in a wide spectrum noise is masked only by the components of the noise that are near it. The audiograms show that (at least for low SPLs of the masker) components added to a band of critical width do not add to the masking of a tone at the peak of the audiogram, but the subtraction of components from a masking stimulus of critical width lowers the threshold by an amount equal to the power contained in the subtracted components. The frequency components falling within a critical band are sumed to produce an effect dependent on total power. The dip that appears in the masked audiograms produced by two pure tones can be interpreted to mean that the critical bandwidth is the frequency interval that must separate two tones in order for them to produce "unsummed" or separate effects. Masking can now be seen as intensity discrimination: The power of the signal falling within the critical band will be integrated with the power in the noise. The masked threshold of a pure tone is then the intensity required to produce a just noticeable difference in the intensity of the critical band of noise with which the tone is summed. It should be noted that the duration of the tone bursts (167 msec in this case) has some effect upon the results and that this effect is frequency dependent.

An interesting result form the point of view of constructing models of masking was that, independent of the level of the masking noise, and for subcritical maskers, the masked audiograms produced were the same a long as the noise was of equal power and of the same centre frequency (i.e. they were independent of the actual bandwidth of the masker).

Using the above results Greenwood also concluded that the critical bandwidth was that frequency interval over which the cochlea sums power and that critical bands represent equal distances on the BM, one critical band corresponding to one milimeter [13]. Critical bandwidth is equivalent to the first derivative of the characteristic frequency position function along the BM.

The concept of critical bands is a recurrent one in all areas of psychoacoustics. Greenwood [11] showed that masking patterns of subracritical stimuli were different than those of subcritical stimuli. Since masking is essentially a reduction in loudness similar effects should be expected for the loudness of wideband stimuli. In [10] a detailed study of loudness of groups of tones and bands of noise was undertaken. In particular it concerned how the loudness of a group of tones depends on the spacing of the tones in the complex and how the loudness of a band of noise of constant SPL depends on the width of the band. For a four tone complex the spacing of the (equally intense) components was changed and its loudness compared with a variable comparison signal of known loudness Figure [2.3-17] shows the effect on loudness produced by changing the overall spacing Δf of four tones spaced around the frequencies 500,1000 and 2000 Hz.. The overall spacing corresponds to the frequency difference between the highest and lowest tone. The different symbols (T and C) correspond to two different procedures used: T = single tone adjusted to match the loudness of the complex and C = complex adjusted to match the loudness of the tone. The lines through the data show clearly a knee in the response when the overall spacing Δf exceeds a critical value. The results agree with the hypothesis that within a critical band the loudness is independent of the spacing of the tones. When the overall spacing Δf exceeds a critical value the loudness increases. This phenomenon was observed at all levels of the SPL of the complex where the knee occurred at the same overall spacing independent of level, apart from at very low SPLs (~20 dB) where no knee was observed. They concluded that at these low levels the subjects had changed their loudness criterion for the comparisons. They found that the subjects instead of "integrating" the loudness of the components, they tended to judge the total loudness to be equal to that of a single component in the complex. It seems that for low levels and wide spacings "loudness integration" may break down.

Another interesting result was obtained when the overall spacing was kept constant but the relative spacing of the tones was changed. They found that uniform spacing produces greater loudness than nonuniform spacing. The difference could be as high as 3 dB. This could be of some significance to RELP coders (see chapter on speech coding) where band folding or band translation is performed for regeneration. At the band edges there is a discontinuity in the harmonic structure, but within each band the harmonics are regularly spaced. It may be that the loudness of a "discontinuity" band is lower than the loudness of a "continuous" band a fact that would enhance the quality of the dedcoded speech. Similar experiments with bands of noise revealed that the loudness of the noise of constant SPL is invariant with bandwidth provided that the bandwidth is smaller than a critical band. [2.3-18]. When the bandwidth of the noise is increased its loudness increases although its SPL is held constant.

This, again is relevant to speech coding: It is well known that a flat (white) noise is perceptually louder than a noise which is shaped according to the shape of the speech signal. Although this effect is at least partly due to the more efficient masking of the noise by the speech signal, even if the coding noise was presented on its own i.e. the speech signal removed, its loudness when flat would be higher than when it is shaped. this is because shaped noise is confined within areas of small bandwidth (the formants) which would be subcritical instead of being summed over the whole spectrum. A similar argument applies when the noise concentration is performed in frequency domain coders such as subband coding.

In [14] Scharf performed similar experiments under masking. In his experiments he did not use white (flat spectrum) noise but uniform masking noise. This was produced by passing the output of a white-noise generator through a filter whose power attenuation increased proportionately with frequency above about 500 Hz. Such noise raises the thresholds for most of the audible frequencies to approximately the same level. Note that from the experiments of Greenwood [11], noise at constant SPL within a critical band produced the same amount of masking irrespective of bandwidth. If white noise was used with a flat spectrum, the SPL falling within each critical band would increase with frequency simply because the bandwidth of a critical band increases with frequency. Therefore the amount of masking introduced by the noise would likewise increase with frequency. This can be kept constant if the spectrum of the noise decreases in accordance with the increase in critical bandwidth, so that a constant SPL is achieved within each critical band.

This again reflects upon noise shaping in speech coding. The masking effect of noise is reduced if its spectrum is made to follow the (lowpass) spectrum of speech.

Scharf found that when noise was not excessive compared with the tone complexes his results were similar to the previous study [10].

It is important to note that the increase in loudness as the bandwidth is increased at a constant SPL is closely related to the power law function of loudness. If one accepts that the loudness of a complex sound is a fixed function of the sum of the loudness of the component critical bands, an increase in loudness with bandwidth of constant SPL will only be present if changes in loudness are smaller than changes in intensity. Take the case of a band of noise. If its bandwidth is one critical band, then

$$N_c = \left(\frac{I}{Io}\right)^{0.3} \qquad 2.3\text{-}6$$

where N_c is the loudness and I the intensity of the noise. If now the noise is split into two critical bands with the overall intensity the same as before the new loudness is given by

$$N_{2c} = \left(\frac{I}{2Io}\right)^{0.3} + \left(\frac{I}{2Io}\right)^{0.3} = \left(\frac{I}{Io}\right)^{0.3} . \quad \frac{2}{(2)^{0.3}}$$

2.3-7a

hence

$$N_{2c} = N_c \ . \ \frac{2}{(2)^{0.3}} \qquad 2.3\text{-}7b$$

therefore the loudness is increased because $2>(2)^{0.3}$ or $0.3<1$. The increase in loudness can be attributed to the compressing nonlinearity of transformation from intensity to loudness.

Since the shape of the loudness versus intensity curve changes near threshold (approaches linearity near threshold) the effect of loudness increase would be expected to be less prominent. This is indeed the case as was found in [14].

2.3.3 Partial Masking

In the experiments of Greenwood [11] as in all studies aimed at obtaining audiograms or masking curves [6, 7, 8, 9] the interest is centered upon complete masking or determination of thresholds. In complete masking, one sound interferes with a second sound so much so as to make it inaudible. In loudness calculation partial masking is also important. This is the case where the first sound while not completely masking the second sound, reduces its loudness. Experiments on partial masking are reported by Scharf in [15]. His aim was to show how the partial masking of a pure tone by a narrow band of noise depends upon the respective SPLs of the two sounds and upon their frequency separation. He used the same masking noise, one critical band wide and centered at 980 Hz throughout. The tone to be masked was at one of five frequencies, 690, 830, 980, 1155 and 1355 Hz (spaced one critical band apart). The SPL of the masked tone was held constant at 25, 45, 65 or 85 dB and its loudness measured as a function of the SPL of the masking noise. Figures [2.3-19, a-e] show how the loudness of the masked tone decreased as the SPL of the masking noise increased. Each figure represents the curves for a different frequency. The parameter on the curves is the SPL of the masked tone while the ordinate gives the SPL of the comparison tone (the tone that was found to have the same loudness in the quiet as that of the masked tone).

The curves are horizontal over the range of noise levels at which the noise was too soft to reduce the loudness of the masked tone. On these portions of the curves, the masked and comparison tones always have the same SPL, since being of the same frequency they must be equally loud when equally intense. The curves depart form their horizontal course to follow the experimental points down to the masked thresholds, represented by the symbol T. The symbol T has the value on the ordinate of the mean absolute threshold for the comparison tone in the quiet; its value on the abscissa is the minimum noise level required to mask completely the tone when set at its parametric SPL. Generally the threshold measurements and the loudness measurements are in good agreement in that they all fall on the same smooth curves. The same auditory process appears to determine the outcome of both types of measurements.

Note especially the rapid growth of the loudness of the masked tone as the noise is reduced to the level of complete masking. Although the results from the threshold values show that high frequency tones are completely masked by less intense noise than the low frequency tones (as expected), the tones lying above the frequency range figure [2.3-19d,e] of the noise grow more rapidly in loudness as the noise level is reduced than do tones lying below (fig. 2.3-19a,b).

This difference can be accounted for by the same asymmetrical spread of activity within the auditory system. Consider the schematic representation of these patterns: Figure 2.3-20 shows the idealized spectrum and the assumed neural activity which will be refered to as the excitation pattern for the band of noise (the masker) and also for the tone lying two critical bands above the center frequency of the noise and for the tone lying two critical bands below it. The abscissae are marked in critical band units [12]. The ordinate gives the relative amplitude for the spectra and the level of (neutral) excitation in dBs for the excitation patterns. The shape and level of the pattern were determined from a masking curve produced by a narrow band of noise. If effects due to beats and harmonics are excluded, the masking pattern for the tone is similar to that of a narrow noise band.

These patterns change little over the limited range of frequencies shown here [11]. The derivation of excitation patterns (neural activity) from masking patterns is justified by recent experiments [7,8].

To explain partial masking the assumption is made that this occurs only when the excitation patterns overlap and further that within the areas of overlap the pattern at the higher excitation level suppresses the pattern at the lower level. The size of the shaded "suppressed" area indicates the relative degree to which the tone is masked - the larger the shaded area the greater the partial masking. A comparison of the shaded areas for the low frequency with those for the high frequency tone shows clearly that the partial masking of the high frequency tone decreases more rapidly as the noise intensity decreases than does the partial masking of the low frequency tone.

These differences result from the skewness of the excitation patterns. The implication here is that for partial masking, as the noise level is increased, the noise begins to partially mask the low frequency tones at lower levels than the high frequency tones.

In [16] by direct analogy with threshold audiograms Scharf produced partial masking audiograms. His results are shown in figures (2.3-21, a-e). The abscissa gives the frequency of the signal. The ordinate gives the amount of masking, defined as the amount by which the SPL of the signal had to be increased, owing to the presence of noise, in order to stay at a given criterion level. In complete masking the criterion was threshold. In partial masking, the criterion was the loudness level of the tone in quiet, which is the parameter on each curve, so that the curves are similar to equal-loudness contours. The amount of masking is always the difference between the SPL of the tone in noise and the equally loud tone in quiet. The top contour is the threshold curves. The rectangle with diagonal lines centred on 1000 Hz indicates the frequency limits of the masker. It can be seen that the threshold curve and the first (low loudness) curve are parallel. This is not a trivial result since for the second case the curve was obtained through loudness matching whilst the threshold curve involved no such procedure. As the loudness level increased and the masking reduced, the asymmetry of masking diminished and the patterns become fairly symmetrical or even skewed toward the low frequencies at the highest loudness levels, as was also found in the previous study [15].

The asymmetry in partial masking increased with the level of the masker in the same way as the threshold curves. A reduction in the bandwidth of the masking noise narrowed the spread of masking. This is at variance with the assumption that subcritical bands of noise at equal SPLs produce the same amount of masking and loudness. This may be due to the time fluctuations of the noise band (the broader band permits faster fluctuations than the narrower band).The subject may have been able to hear the signal in the time "valleys" of the noise. The effects of time fluctuation on masking by noise will be addressed again later. Before we move into time effects mention should be given to some recent experiments concerned with the masking of noise by tone. Hellman [17] addressed the problem concerning the partial masking of noise by a tone. Amounts of partial and complete masking produced by the tone on the noise were obtained for subcritical, critical and supracritical bands of noise. In the first instance a 1KHz tone masked a narrow band noise 925-1,080 Hz wide, an octave band noise 600-1,200 Hz wide and a broadband noise 75-9,600 Hz wide. Masked and unmasked noise bands were matched in loudness, to produce the measurements. Median values were used (these matched geometric means) for the determination of the average SPLs of the noise. The results are shown in figures [2.3-22, 2.3-23]. The data show the SPLs required for equal loudness between a band of noise in the presence of a pure tone and the same noise in the absence of the tone. Figures [2.3-22 a,b] show the masking effect of the tone on broad-band and octave band noise respectively. The effect of the tone on the narrow band noise is shown in figure [2.3-23]. The parameter on the curves is the SPL of the masking tone. Over the range of intensities used, the loudness-matching functions are power functions. For the wide band signal the slope (exponent of the power function) is shown to vary form about 1.1 for a masking tone at 80 dB. SPL to about 1.35 for a tone at 100 dB SPL. It can be seen that when the tone is below 90 dB SPL it has practically no masking effect on the wideband noise. The masking effect increases when the signal bandwidth is decreased to an octave. The slopes vary from about 1.25 for the 80 dB SPL tone to about 1.45 for tone at 100 dB SPL.

A sharp increase in the masking effect of the tone is produced when the noise bandwidth is further reduced to 155 Hz. The slopes increase from about 2.25 for a masking tone at 60 dB SPL to about 2.6 for a tone at 90 dB SPL. This steepening of the power functions can be attributed to the fact that the spread of excitation of the tone and noise is more nearly the same. Signal to noise ratios for the narrow (925-1080 Hz) signal at threshold are about -20 dB when the masking tone is at 60 dB SPL increasing to -30 dB when the tone is at 90 dB SPL. By contrast when the role of noise as signal and tone as a masker are reversed, the signal to noise ratio at the tone threshold is of the order of -4 dB.

To a good approximation, the functions in figures [2.3-22, 2.3-23] show that the tone no longer masks the noise when the energy of the tone and the energy in the 155 Hz noise band surrounding the tone are equal.

This means that for this noise band, partial masking ceases when the tone and the noise are at the same overall SPL, while for the octave band the overall SPL of the noise is about 6 dB above that of the tone and for the broad band noise it is about 18 dB above that of the tone before masking ceases. Hence masking ceases when the masked and masking stimuli are equally intense within the effective area of masking. This area is the critical band for the tone. The ear filters out the energy contributed by frequencies of noise outside the critical band for the tone. Note that along the steep portion of the masked loudness function mutual masking between tone and noise occurs. The asymmetry in (complete) masking found here (around 20 dB) between tone as masker upon noise and noise as a masker upon tone could be due to the rate of fluctuation in the time structure of the noise. Indeed, the asymmetry seems to disappear when the tone is frequency modulated at a rate of 25/Hz [20].

In [18,19] the same problem of using a tone as a masker was addressed. An explanation to the asymmetry of masking was provided which is essentially the same as that from the previous authors although numerical values for the asymetry were successfully predicted. Their argument is as follows: The noise power N as measured by the human ear has an uncertainty ΔN given approximately by $\Delta N = N/\sqrt{TW}$ where T is the integration time of the ear and W the analysis bandwidth. With T = 0.1s and W = 160 Hz (at 1 kHz) $\Delta N = N/4$ or -6 dB. Thus a tone whose level is 6 dB below the noise level would be hard to detect since the measured noise fluctuations are equally large (in reality, a 1-kHz tone is inaudible at -3dB). By contrast a critical band noise added to a pure tone is perceived at much lower levels because, even minute deviations from a pure tone are easily detected by the ear. The added noise will cause the instantaneous amplitude and frequency of the tone to fluctuate with time instead of being rigidly constant. An amplitude modulated 1kHz tone at 60 dB SPL will be perceived as different from a pure tone if the modulation index exceeds about 8% independent of the modulation frequency. A modulation index of 8% will be caused by an additive noise of level -22dB below the tone, which is a value close to the measurements obtained from experiments in

[18]. These additional experiments on the effects of tone as a masker are presented below: for the results shown in fig. 2.3-24 a continuous tone at an SPL of 80 dB was used to mask a noise burst fixed in frequency at 1kHz (centre) with a bandwidth of one critical band. The intensity of the noise burst was varied so that it was just above threshold as the frequency of the masker tone was varied. Note that for a tone of frequency of 1 kHz the noise intensity at threshold is 24 dB below the tone intensity.

The shape of the audiogram has a "reversed" character since the maskee (probe) location was fixed as the masker was varied. The opposite is true for conventional (masked) audiograms. Partial masking curves were then obtained [figure 2.3-25]. Here again, the 80 dB SPL noise burst is used as a fixed probe of bandwidth of one critical band and centered at 1 kHz. The insert on each plot is the frequency of the masker tone. These results are similar to the ones obtained by Hellman [17]. It should be noted though that these curves were obtained from only one subject. The usual practise is to use at least 3 subjects but usually 5 and average the results in a statistically meaningful way.

2.3.4 Temporal effects

2.3.4.1 Temporal Summation

The (masked) threshold as well as the loudness of a sound are independent of its duration when the duration exceeds about 500ms. However for durations less than about 20 ms the threshold increases and the loudness decreases as the duration of the sound (e.g. tone bust) is decreased. Over a reasonable range of durations the ear appears to integrate the energy of the stimulus over time. (to a first approximation) [21-25]. To decouple bandwidth effects from the studies care is taken in experiments of temporal integration to confine the spectrum of the signal to within a critical band. Note that this restriction limits the minimum duration to be used since reducing the signal further would necessarily imply an increase in its effective bandwidth. Using tone bursts as test-sounds, the results can be represented by a simple graph, which holds for the masked threshold as well as for the threshold in quiet.

In the graph [figure 2.3-26] the SPL, L_T (T_T) of a test tone is shown, versus its duration T_T. The asymptote represents the threshold of a long duration (>500 msec) tone $L_T(\infty)$. The vertical axis units measure ΔL_T where $\Delta L_T = L_T(T_T) - L_T(\infty)$. This is the difference between the threshold of the tone of long duration and the threshold of a tone of duration T_T for the same masker conditions. It can be seen that for long durations (T_T > 500msec) the "normal" threshold is reached. For durations shorter than about 200 msec, ΔL_T rises 10 dB if the duration is shortened by a factor of 10. This holds for the masked threshold as well as for the threshold in quiet and for both tone bursts and narrow bands of voice. The general behaviour of the threshold can be described by the following equation

$$\Delta L_T = L_T(T_T) - L_T(\infty) = 10\log_{10} \frac{1}{1\text{-EXP}\left(\frac{-T_T}{0.2}\right)} \text{ dB} \qquad 2.3\text{-}8$$

when time is measured in seconds.

This simplifies to

$$\Delta L_T = 10 \log_{10} \frac{0.2}{T_T} \text{ dB} \qquad 2.3\text{-}9$$

for short durations. (from the series expansion of e^x).

The effect of duration upon loudness is very similar figure [2.3-27]. T_T is, again, the duration of the tone burst (or narrow band noise) and ΔL_N is the difference in loudness level between a long duration signal and its counterpart, of duration T_T, for the same masker condition. It can be seen that the loudness level is constant for long durations (above 200 msec). It decreases for shorter signals about 10 phon per factor of 10 of shortening. For such an approximation, 100 msec is the limiting duration as shown in figure [2.3-27], by the intersection of the asymptotes.

The departure of the curve from the asymptote at very short durations (<5 msec) is due to the spread of the signal's bandwidth outside the critical band. Besides the temporal effect, a spectral effect comes into play.

2.3.4.2 Transient effects in masking: Backward Masking, Forward Masking and overshoot

In the previous section the effects of varying the duration of the maskee in relation to its masked threshold and loudness (i.e. effects of complete and partial masking) were investigated. In the above experiments a continuous masker was used. Many other temporal effects in masking can be observed and measured, if now, the masker häs a strong time structure. To obtain these time-dependent masking patterns, a very short duration (but contained within a critical band) tone is used as a probe. This is then moved within the time evolution of the masker (which for the time being will be considered to be noise) and the threshold for each relative position of the probe is measured. A distinction must be made between narrow band and wideband maskers. We have already seen that tones and critical band noises produce different masking patterns.

This seems to be contrary to the conceptual idea of a critical band auditory filter which implies that the two stimuli should give rise to the same masking pattern when at the same SPL.

It has already been indicated that this may be due to the slow time fluctuations that a narrow bandwidth imposes upon the time structure of the masker. The subject can then hear the probe tone in the "valleys" of the noise. When the noise bandwidth is increased the corresponding fluctuations become faster and it seems that the subject is not able to follow them any more. Fastl [26] performed some masking experiments with both critical and very narrow band noises. He used high frequency ranges in his experiments (8.5 kHz) since at these frequencies the critical bandwidth is large (1800 Hz) which allows large variations in the bandwidth of subcritical maskers. His results generally agree with Greenwood's [11] in that the tone masker is not as effective as a critical band marker. Fastl's wider critical bandwidth enabled him to use very narrow band noise as compared with critical bandwidth (1/18 and 1/180 of critical bandwidth). For these maskers the masking patterns were essentially the same as that of the tone. The effects of narrow and wideband maskers will therefore be considered separately.

The set up used for time varying maskers is usually as shown in figure [2.3-28]. Bursts of noise of equal width T_M and presented at equal intervals T_P are used as maskers. A probe tone of duration T_T is used to measure masked thresholds. Δt is the interval from the onset of the noise burst to the onset of the tone burst. Note that Δt can exceed T_M since masking effects are present even after the cessation of the masker (forward masking, post masking) or Δt can be negative since the masker has some effect on the threshold of the tone before the masker is presented (backward masking, pre-masking). This may seem paradoxical but it can be explained by the finite analysis window of the auditory channel and the appropriate delay associated with it. The general shape of the masked threshold around one of the noise bursts can be seen in figure [2.3-29].

Transient effects occur also in the case of $0<\Delta t<T_M$ i.e. simultaneous masking [21, 27. 28]. If one ignores effects smaller than 2 or 3 dB then the threshold of a short signal pulse, masked by a burst of the masker, shows a transient with an overshoot when considered as a function of the delay time between the onset of the masker and the onset of the signal, provided that the signal and masker have different shaped frequency spectra. This overshoot does not show up if signal and masker have the same or similar frequency spectra. The similarity must hold around the frequency of the (maskee) signal. The more the two spectra differ from each other, the more overshoot one gets.

No overshoot occurs if signal and masker both have narrow frequency spectra. This holds even if masker and signal are located at different centre frequencies. The amount of the overshoot depends very little on the level of the masker but falls to zero near absolute threshold. The overshot is also influenced by the size of the time gap between the bursts of the masker, and in the case of a continuous background masker present, it also depends on the level difference between the gated and continuous masker.

The overshoot disappears for signal (probe) durations larger than 10 msec and grows to as much as 15 dB for durations of 2msec. The effect lasts for up to 100 msec where the steady-state condition is reached (provided the masker duration T_M is long enough).When the masker burst is short, the thresholds of short pulses with small delay between onset of masker and onset of signal are not affected by the amount by which the masker outlasts the signal, except when the duration of the masker becomes smaller than 10 msec.

Fastl [30-32] performed a very extensive series of experiments to determine the parameters that affect the magnitude and extent of the above transient events. Three different types of masker were used, broadband, critical and sinusoidal. His results were very illuminating and a summary is presented in the following sections.

2.3.4.3 Transient Masking effects: Broadband Noise Masker

In [31] Fastl used masker impulses out of uniform masking noise (a noise that produces the same degree of masking at every frequency). Impulses with Gaussian rize and decay cut out of pure tones were used as test signals (probes). The durations of masker and signal were chosen in such a way that their reciprocal was never greater than the critical bandwidth at their corresponding frequencies. His aim was to produce a complete map of the masking effects of the uniform masking noise burst. He therefore studied all of backward masking, simultaneous masking and forward masking. His results are summarised below:

2.3.4.3a Simultaneous Masking:

Since short test tone impulses must be used to gain insight into the fine structure of temporal masking patterns, its first experiment was to study the effect of shortening the test time on the shape of the patterns. At long test durations (500 ms) the masked threshold was found to be independent of test tone frequency as was expected with a uniform masking noise. At short test tones (10 ms) however the masked threshold decreased with increasing frequency. The difference was at most 10 dB. He then performed some experiments to study the overshoot effect. He found that the magnitude of the overshoot increases as the delay between onset times is decreased reaching 20 dB in some cases and that the effect is more pronounced with short test tone impulses. The steady state condition is reached for impulse duration times of at least 200 ms when no overshoot is present. Hence it seems that the time between the onset of the masker and the end of the tone is the relevant parameter in terms of onset delay.

In terms of test tone frequency, the data showed an increasing overshoot with increasing test tone frequency. The maximum difference was about 9 dB but the effect was not consistent amongst individual subjects. As far as the masker level is concerned, the overshoot effect is distinct provided the masked threshold of the test impulse lies more than 10 dB above its threshold in quiet. Finally with a test tone of 1 ms duration, the masker duration was varied through the values 2, 5, 10, 20, 50, 100, 200 and 500 ms. The onset delay was 1 ms. At all masker durations the masked thresholds were the same within 3 dB.(comparable to the accuracy of measurements). This means that at the start of a masker impulse the thresholds of simultaneously presented test tones are nearly the same for masker impulses of extremely different durations despite the fact that these masking impulses differ considerably in loudness (because of loudness integration).

2.3.4.3b Forward masking:

A typical forward masking curve is shown in figure [2.3-30]. The level L_T of the just audible test tone impulse is plotted as a function of the delay time T_v. Test tone impulses with frequency f_T = 8 kHz and duration T_T = 1 ms were used. Masker impulses with duration T_M = 500 ms were cut out of a uniform masking noise with level L_M = 60 dB and a bandwidth Δf_M = 16 kHz. The arrow at the left ordinate marks the threshold in quiet of the test tone impulse. Temporal and spectral relations between masker and test tone are indicated by inserts. It can be seen that the forward masking function can be approximated by a straight line, when using two logarithmic scales, for loudness level and time. The test tone level at short delay times t_v in a forward masking paradigm is nearly equal to the test tone level at long delay times in a simultaneous masking paradigm (top left of figure (2.3-30). It was found that at any (fixed) delay time the masking effectiveness of a preceding masker with respect to short test tone impulses hardly depends on their durations. This is not to be taken to mean that the threshold was the same at all durations, but rather that the level difference between (forward) masked threshold and threshold in quiet for the same duration test tone is almost independent of duration.

When the masker level was now varied, the results showed that the forward masking function reaches the threshold in quiet at a fixed delay time, irrespective of masker level. This means that for high masker level steeper forward masking functions show up than for lower masker levels. This bears some effect on the masking pattern of an uniform masking noise impulse as a function of frequency. Consider a medium delay, say 12 ms. Since the value of the threshold in quiet is not constant but varies with frequency, having a minimum at around 3 kHz, the forward masking curves will be steeper around 3 kHz. Therefore at the fixed delay the threshold at 3 kHz will be lower than at the other frequencies and the forward masking pattern resembles to some degree the frequency dependence of the threshold in quiet.

Finally the effect of masker duration on forward masking was discussed. The forward masking effect was found to be more pronounced for long masker impulses than for short masker impulses. For masker durations longer than 100 ms, no further increase in masking effectiveness was noticed. Whereas in forward masking experiments, shorter maskers elicit lower masked thresholds, the simultaneously masked threshold at short delay times was found to be nearly independent of masker duration. Thus at short masker durations, near the end of the masker impulse very steep forward masking functions have to be expected.

2.3.4.3c Backward masking:

A typical backward masking function is depicted in figure [2,3-31]. The test tone level L_T is plotted as a function of the delay time Δt. Since Δt is measured from the start of the masker impulse to the start of the tone impulse, only negative values of Δt occur for backward masking. Both spectral and temporal relations between masker and test tone are indicated in figure 2.3-31. The arrow at the left ordinate marks the threshold in quiet of the test tone impulse with frequency f_T = 8 kHz and duration t_T = 1 ms. This short test tone duration was chosen in order to reach a sufficient temporal resolution within the backward masking function. Threshold discrimination in backward masking was found much more difficult than either simultaneous or forward masking. Experimental variability as high as 10 dB was not uncommon. From figure 2.3-31 the backward masking function shows a very steep slope near the start of the masker impulse (25 dB in 3 ms). When Δt reaches -20ms the masked threshold is already very near the threshold in quiet. Masking functions were obtained for test tone duration T_T, test tone frequency f_T masker level L_M and masker duration T_M. The results can be summarized as follows: Backward masking is found for short test tone impulses, presented up to about 200 ms before the start of the masker impulse and might depend to some extent on test tone duration. At medium delay times, the backward masking pattern of an uniform masking noise impulse resembles the frequency dependence of the threshold in quiet. No extremely nonlinear relation between masker level and test tone level shows up in backward masking. At long masker impulses, more backward masking occurs than at short maskers.

2.3.4.3d Transient masking pattern

To give a survey about relations between backward simultaneous and forward masking patterns in fig. [2.3-32] a transient masking pattern is depicted. The test tone level L_T at threshold is plotted as a function of both critical band rate Z (the frequency scale compressed such that critical bands are shown to be of equal width) and time t. The masker impulse was cut out of uniform masking noise, starting at t = 0 ms and terminating at t = 300 ms with an SPL of L_M = 60 dB (Δf_M = 16 kHz). The test tone impulses had a duration of T_T = 10 ms with a rise time Tr_G = 2 ms. Negative values of t refer to backward masking effects while t = 300 ms denotes the end of simultaneous masking and the beginning of forward masking. At t = -20 ms and t = 400 ms the respective backward or forward masking pattern represents the pattern of the threshold in quiet within 5 dB. This can be taken to be the start and the termination of the pattern respectively. This pattern can be assumed to show the spectral and temporal representation (of the masker) in the ear.

Using the above data as an atlas of temporal masking effects produced by a single (of duration larger than 500 ms) broad band masker impulse, thresholds of test tone impulses masked by broad band noise of various temporal structures can be estimated.

In figure [2.3-33] a masking noise burst is represented by hatched areas. The burst consists of single impulses with duration 1 ms separated by gaps of duration 2 ms. Using a 1 ms test tone impulse, the threshold L_T is plotted (dots). Calculated values are shown by crosses. The values were calculated using the data from before. Good agreement obtains; differences are no more than 5 dB. An important conclusion is that in temporal masking functions, the ear cannot resolve gaps of 2 ms (but note that the test tone impulse used was of comparable time length).

A similar experiment using a masker burst consisting of impulses with duration 2 ms separated by gaps of 20 ms is shown in figure [2.3-34]. Again calculated and measured data are in good agreement. Note the fast decay of forward masking tone due to the short duration of the masker. Finally the effect of varying the gap between the impulses was studied [fig. 2.3-35]. Longer decay of forward masking can be observed corresponding to longer duration maskers. Measurements agree well with calculations except in regions where backward and forward masking interact. In these cases the threshold is elevated more that the predicted values suggest. Another result that can be observed is that a "deep valley" preceding a second respective masker impulse (due to the fact that the preceding first impulse was sufficiently far away for masking to decay considerably) gives rise to a high threshold value (large overshoot) whereas a "shallow valley" (due to the preceding impulse being near enough in time) leads to a lower threshold (small overshoot).

2.3.4.4 Transient effects in masking: Critical band noise

A similar set of experiments was carried out in [32]. In these experiments masker impulses were cut out of critical band noise instead of a broad band noise. A masker centre frequency of 8.5 kHz was used which would enable impulses of sufficiently short duration (1 ms) to be contained within a critical band.

The results can be summarised as follows:

2.3.4.4a Simultaneous masking:

The shape of the masking pattern of a continuous critical band noise masker is independent of test tone duration. Thresholds of test tone impulses, masked simultaneously by a critical band noise masker impulse, dependent on delay time as follows: test tones centred in the critical band noise are almost not affected by variations of delay time. Test tones at the lower slope of the masking pattern show decreasing threshold values with increasing delay time. At the upper slope of the masking pattern, this decrease is observed only at low masker levels. At the start of short and long critical band noise masker impulses, respectively, almost the same threshold values show up. With increasing masker level the pattern exhibits a "nonlinearity of its upper slope" both at short and long delay times.

2.3.4.4b Forward masking:

From experiments on forward masking the following have been concluded:

Forward masking patterns of a critical band noise masker impulse èxhibit steeper slopes than corresponding simultaneous masking patterns. The maximum of the forward masking pattern shows up at a higher frequency for medium delay time than for short delay time. With increasing masker level the masked threshold of test tones at the upper slope of the forward masking pattern of a critical band noise masker increases more than proportional. At a fixed delay time, short critical band noise masker impulses elicit less forward masking than long maskers.

2.3.4.4c Backward masking:

Again, the backward masking data proved more difficult to collect, since the reproducibility of the experiments was distinctly inferior to the reproducibility found in both simultaneous and forward masking. Conspicuous results on backward masking are summarized below: Backward masking functions of critical band noise masker impulses show, near the start of the masker impulse, a horizontal course. Backward masking patterns of critical band noise masker impulses at short delay times are therefore quite similar to corresponding simultaneous masking patterns. The "nonlinearity of the upper slope of the masking pattern" exists in backward masking too.

At a constant delay time, very short critical band noise masker impulses elicit less backward masking than longer impulses.

2.3.4.4d Transient masking pattern:

The transient masking pattern of a critical band masker is shown in figure [2.3-36]. The level of the masker was 70 dB its centre frequency was 8.5 kHz, its bandwidth 1800 Hz and its total duration 500 ms. The test tone impulse was of 1 ms duration. Regarding figure [2.3-36], the similarity of neighbouring backward and simultaneous masking patterns as well as the masked differences of simultaneous and forward masking patterns are obvious. Forward masking patterns exhibit steep slopes towards both low and high frequencies whereas simultaneous and backward masking curves show steep slopes only towards low frequencies. The transient masking pattern represents a common starting point for the description of hearing sensations of both static and dynamic character.

When more than one critical band noise masker impulse is used, thresholds of test tone impulses masked by a critical band of a strong temporal structure can be estimated. The results are shown in figures 2.3-37 and 2.3-38. Circles connected by solid lines mark measured thresholds, dashed lines represent estimated masking functions. The estimations were based on the data collected from the single impulse masker. Estimates can be seen to agree fairly well with the measured data over the regions where forward and backward masking do not interact.

2.3.4.5 Transient Masking Effects: Sinusoidal Masker

In [30] similar experiments were carried out, using a masking impulse cut out of a sinusoid. The masker frequency was 8.5 kHz, the SPL L_M = 70 dB and the duration t_M = 200 ms. The test tone impulses were of duration T_T = 2 ms. The transient masking pattern obtained is shown in figure [2.3-39]. Although results are somewhat similar to the results of a critical band masker two main differences can be noticed: (a) The transient masking pattern of the sinusoidal masker sows steeper slopes than the pattern of the critical-band masker. (b) The masking produced by a sinusoidal masker is inferior to that produced by a critical band masker. The whole transient masking pattern of the sinusoid lies 10 dB below that of the critical band masker (the shapes at both ends of the transient masking pattern are influenced by the shape of the threshold in quiet).

In physiological experiments (e.g. to obtain tuning curves) sinusoidal sounds are used. It is not therefore known whether differences in masking patterns between sinusoids and critical band maskers appear also in physiological data.

2.3.4.6 Statistical changes:

To investigate to which extent the ear is able to follow random temporal changes of SPL, Zwicker and Schütte [33] used a computer generated pseudo random noise with a large repetition period (~1 sec). With this type of noise repeatable time functions can be produced with little rhythm.

The temporal characteristics of such a noise can be changed by altering the bandwidth: for large bandwidths the changes in amplitude and frequency occur very quickly. For very small bandwidths, the amplitude changes appear so slowly (frequency changes can be neglected in that case) that a quasi steady state condition is reached. The masking patterns would therefore be expected to follow the changes of the sound pressure for very small bandwidths whereas for large bandwidths the masking pattern may not be able to follow the quick changes and may therefore average them in a way so that another steady state condition is reached.

A short tone-signal was used to map the masking patterns of the artificial noise. Its duration was 2 ms. Care was taken so that its bandwidth did not exceed a critical bandwidth. The artificial noise was filtered by narrow band filters of different widths, mostly at a centre frequency of 4 kHz. The masked threshold L_T measured at delay times Δt within a certain range of the time evolution of the masker is compared with the instantaneous sound pressure level L_{NBN}. The rate of the amplitude changes of the narrow band noise is restricted by the limited bandwidth. The short signal represented by its masked threshold L_T is able to follow the slow changes quite well (figure 2.3-40) while it cannot follow the quick changes of the masker at large bandwidths (figure 2.3-41). A statistical analysis of how well the masked pattern follows the instantaneous envelope of the noise was performed for different noise bandwidths. Let Y represent the measured threshold levels L_T and X the instantaneous levels L_{NBN} (at the same delay time Δt). Two statistical measures were used, first the linear regression factor m defined as

$$m = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \qquad 2.3\text{-}10$$

note that m given as above is the least squares estimate for m in Y = mX + b.

The second factor used was the correlation coefficient r given by

$$r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \overline{X})^2 \sum_{i=1}^{n} (Y_i - \overline{Y})^2}} \qquad 2.3\text{-}11$$

Note that whilst r could be large (good correlation) m can be far from one (if the amount of change of L_T is larger or smaller than L_{NBN}).

The measurements were made at a center frequency of 4 KHz. Bandwidths of 12, 32, 100, 400 and 1000 Hz were used. (The critical bandwidth is about 1 kHz at that frequency). The results are shown in figure (2.3-42) as a function of bandwidth.

For small bandwidths m and r are very close to 1. This means that L_T and L_{NBN} change in the same way for the same amount. At a bandwidth of 100 Hz the correlation coefficient is still 0.85 indicating that L_T and L_{NBN}

have a similar temporal pattern but the extent of the changes of L_T start to diminish (m = 0.5). For Δf = 400 Hz and greater, both r and m are small, indicating that the threshold L_T no longer follows the temporal variations of the envelope.

In the above experiment the centre frequency of masker and signal were the same. For signal frequencies below the masker center-frequency the behaviour is similar to the one in the above experiment. For signal frequencies above the masker frequency the correlation coefficient behaved as above but m was larger than 1 for some bandwidths, indicating that masking grows faster with masker level there than around the centre frequency. Concluding, the masking pattern was shown to follow the temporal pattern of the instantaneous sound pressure of the stimulating sound almost completely up to a bandwidth Δf of about 100 Hz of the masker. This can be related at a first approximation to a rise time 1/Δf = 10 ms. In [34] H. fastl extended Zwicker's and Schütte's results from masking into loudness measurements. We have seen in previous sections how the masking pattern can be related to neural activity. Loudness is also related to neural activity and masking patterns can be used to derive the loudness of a sound as will be shown in the chapter on psychophysical models.

Firstly Fastl determined the loudness of narrow band noise as a function of the noise's bandwidth. The SPL of the noise was kept constant and all the bandwidths used were less than a critical bandwidth. The noise was centered at 8.5 KHz and bandwidths of 10, 30, 100, 300, 700, 1800 Hz were used. Although previous studies indicated otherwise [10], loudness did change with bandwidth. For bandwidths up to 300 Hz the loudness was the same as that of a sinosoid of equal SPL. Above 300 Hz the loudness of the noise increases to reach levels 10 dB higher than that of a tone at the same centre frequency and SPL, for a critical band noise (1800 Hz). The loudness difference between a sinosoid and an equally intense critical band noise decreased with increased loudness of the sounds.

Taking into account the width of the critical band as a function of frequency [12] this means that up to frequencies of about 2 KHz (critical bandwidth $\Delta f = 300$ Hz) the loudness of sinosoids and equally intense critical (or subcritical) noise bands should be almost the same. Masking patterns between sinosoids and critical band noise are also known to be different [30].

Going back to masking effects, Fastl used the data from Zwicker and Shütte to explain how the effects of non-simultaneous masking could be used to derive masking patterns of narrow noise bands. He attributed the superior masking of wide noise bands as compared to the masking of narrow noise bands of equal SPL to the different frequency of occurrence of envelope maxima in the time course of each noise band. For narrow noise bands the maxima do not occur frequently enough for nonsimultaneous masking to have any noticeable effect. On the contrary, for sufficiently wide band noise, the maxima occur sufficiently frequently to enable forward and backward masking to bridge together the "valleys" between the envelope maxima. In this way the wideband noise produces higher masking thresholds than the equal in SPL narrow band noise. He substantiated his assumptions with some numerical calculations: for a given bandwidth Δf of a narrow band noise, the average number N of envelope maxima per second can be given by

$$N = 0.64\Delta f \qquad 2.3\text{-}12$$

The probability $P(\Delta L)$, for a maximum of the envelope being less than the effective (r.m.s.) SPL of the noise plus a level ΔL can be calculated. Then, the average temporal distance D = 1/N of envelope maxima, lying at least ΔLdB above the effective SPL of the noise can be calculated as follows

$$D = \frac{1}{\Delta f} \; \frac{1}{0.64\,[1-P(\Delta L)]} \qquad 2.3\text{-}13$$

In table (2.3-T1) the average temporal distance D of the envelope maxima, lying at least ΔL = 0, 3, 6, 10 dB above the effective SPL is given for noise bands with bandwidths of Δf = 10, 30, 100, 300, 700, 1800 Hz. From knowledge on nonsimultaneous masking [32] it can be estimated that envelope maxima with temporal distances D>20ms will hardly be weighted by backward and forward masking functions. On the other hand for D shorter than about 6 ms, the weighting functions would more or less bridge the gaps between adjacent envelope maxima. From table 2.3-T1 maxima with a distance D = 2.90 ms lie at least ΔL = 6 dB above the effective SPL of the noise for a noise bandwidth of 1800 Hz whereas for a bandwidth of 300 Hz no appreciable maxima can be found with a separation smaller than about 8 ms. Since the gaps are bridged through nonsimultaneous masking it can be assumed that a 1800 Hz wide masker produces at least 6 dB more masking than an equally intense 300 Hz wide masker. This is true whether the maskee is a short duration signal (<<2 ms) or a long duration signal (>500 ms). Since masking ability and loudness are interrelated through the neural activity they emanate from, a similar difference is expected in terms of their loudness. These calculations agree fairly well with the threshold [33] and loudness [34] measurements.

2.3.4.7 Masking Period Patterns

In the previous sections we studied the effects that noise bursts produce in masking experiments. These studies were extended by "zooming" into the noise bursts themselves using a very short duration probe tone and examining the masking ability of the detailed time envelope fluctuations of the masker and its effect on the masking ability of noise on longer duration signals. To finalise our review of the effects of a masker's time structure on masking we will report on a series of experiments [35-38] which produce masking patterns which are essentially the psychoacoustic equivalent of period histograms. These patterns give a rather detailed account of the phase sensitivity of the auditory system. Phase changes of a partial of a complex tone produce changes in its period's time pattern.

Experiments to investigate the masking effects of a complex tone regarding its temporal fine structure were performed within the masker's period. To avoid averaging across subjects only one single subject would normally take part in such experiments. The test tones were passed through 1/3 octave filters to ensure that the stimulus bandwidth is confined within one critical band. In the first (exploratory) set of experiments [35] the masking patterns of individual tones, a two tone complex, and, DC impulses were investigated. For the first experiments both the masker and maskee were of long duration. In this respect they were different from the more recent experiments we reviewed where the maskee was a very short (~ 2 ms) tone probe.

In the first experiment "octave masking" was investigated. This represents masking patterns of one tone upon a tone of twice the frequency. Tone pairs of 100 & 200 Hz, 200 & 400 Hz, 400 & 800 Hz, 800 & 1600 Hz and 1600 & 3200 Hz were used. The duration was 600 ms for both tones in any one pair. The SPL of a just masked tone (the second in the pair) was obtained for different SPLs of the masker and different phase relationships between the masker and maskee. The result was that the masked threshold of the maskee can change by as much as 15 dB when the only variable that changes is the relative phase of the masker and the maskee tones. This effect was more pronounced for the lower frequency pairs and seemed to disappear for masker frequencies above about 800 Hz.

Experiments with a "two-harmonic-masking" combination as a function of the phase angle of the test tone (masker comprised of two tones, 100 Hz and 200 Hz and maskee at 400 Hz) revealed phase effects of similar magnitude as before whether the relative phase of the masker components was varied or the relative phase between masker and maskee was changed. Other combinations of maskee and maskers were investigated. When the effect of the phase of the test tone on their results was eliminated (by appropriate choice of frequencies) the results clearly indicated that the masking effect depended on the phase angle between the two masker components. In other words masking was found to depend upon the time structure of the sound pressure within one period of the masker. To investigate the effect further small test tone durations were used as in [30-34]. Several complications arise from the characteristics of the stimuli to be investigated and the auditory system itself. Since the test signal must be short in relation to the period of the masker, for a ratio 1/5 and assuming a single masker frequency of 100 Hz, the duration of the test tone becomes 2 ms. Since the test tone must serve as a probe, it must also have a narrow bandwidth where "narrow" means that it must lie within one critical band (any additional narrowing of bandwidth would not be instrumental). For this reason the frequency of the test tone impulse must be greater than about 2 kHz, so that the spectral spread arising from windowing the tone to 2 ms duration is contained within a critical band. The masking functions produced with such "contained" probe tones are called masking period patterns.

The results obtained using a 70 Hz single masker are shown in figure [2.3-43]. Its period was about 14 ms. A test tone impulse with duration 2.5 ms was used as a probe cut out of a tone of frequency 1960Hz. Since the time function of the sound pressure of the masker is repeated after each period T of about 14 ms a train of impulses with a repetition rate fr = 1/T can be used. The pulse train sounds rough but is perceived as a steady state (continuous) signal. The time pattern of the masker as well as the test signal is shown on top of figure [2.3-43].

The value Δt is defined as the time difference between the location of the maximum of the masker and the centre of one of the impulses. Its range is from 0 to T. The (just masked) test signal level L_T is given as a function of the time difference Δt. The level L_1 of the 70 Hz masker is the parameter on the plots. the left horizontal arrow marks the threshold in quiet of the test signal (train of tone impulses). For a masker level of 70 db a high maximum is reached at Δt = (1/8) T while the minimum at Δt = (5/8) T remains near the threshold in quiet: The difference between maximum and minimum values of the threshold reaches 28 dB with only the phase of the masker as a variable! Patterns seem to change shape with changing SPL of the masker. Note that the 2 KHZ probe tone is located at the high frequency end of the "shallow"high frequency slopes of the masking pattern of the masker which are known to exhibit a strong noninear behaviour.

At the right side of figure [2.3-43] masked thresholds L_T produced when a continuous tone is used as a "probe" are indicated. The arrow to the right denotes the threshold in quiet for this frequency. The difference between the level of the two arrows, the one on the left and the one on the right can be attributed to the intensity integration effect which we have seen above

[21-25]. For higher masker level the masked threshold L_T rises quite nonlinearly but there seems to be a correlation between the thresholds measured with continuous test tone and the minimum of the masking period pattern.

More impressive correlation between the tone pattern of the masker within a period and the masking period pattern can be seen in figure [2.3-44]. The masker consisted of two components both at 65 dB SPL of frequencies 112 Hz and 140 Hz (fourth and fifth harmonics of 28 Hz). The frequency of the test tone was 2772 Hz.

When a DC pulse was used as a masker the result shown in figure [2.3-45] was produced. The solid curves belong to a DC- pulse which would be called a "rarefaction pulse" while the dotted curves belong to the "condensation pulses" (the different polarities were obtained by reversing the electrical connection of the earphone).

From these results it seems that the ear is able to listen into or between the "tops and valleys" of the temporal masking pattern produced by the temporal fine structure within the period of the masker. Although the available "time window" that the subject uses is determined by the temporal resolving power of the ear, its location seems to be variable at will so that the best location of the time window can be found, for the ear to look through and extract information. This can be seen from figure [2.3-43] where the threshold for continuous tones is nearly equal to the threshold obtained at the best locations for detection or the "valleys" of the masker. Note that this effect is only true for medium SPLs when the masking period patterns follow the shape of the masker period. When the masking period pattern is nearly flat, no "best locations" exist within the period and the threshold of continuous tones is related to that of the tone impulse through the intensity integration over time (up to 7 dB difference).

In [36] a careful study was undertaken to determine the best parameters for the probe test tone used to obtain masking period patterns. The duration T_T the repetition rate f_P and the frequency f_T of the test tone were varied in order to select the optimal parameters for the best test signal. For these measurements the masker tone with frequency f_M = 110 Hz and SPL at the entrance of the ear canal, L_M = 91 dB was kept constant.

The period T_M of the 110 Hz masker is about 9 ms. Using a test tone frequency of 2640 Hz it was found, as expected, that through a range of test tone durations T_T of 0.5 to 6 ms the threshold difference between maxima and minima of the pattern diminished with increasing T_T, becoming almost zero at 6 ms. It was also found that the difference between maxima and minima was fairly constant from very short durations (0.5ms) up to 3 ms (= 1/3 T_T). Therefore, the maximum difference seems to be obtainable even if the test tone duration is as large as 1/3 of the masker's period.

The repetition rate of the test impulse was not found to have any effect apart from those attributed to intensity integration (i.e an overall shift in level for more frequent repetitions).

Tests with the test-tone frequency as a variable revealed that the shape of the masking period pattern was relatively invariant with frequency as long as f_T was longer than about seven times the masker frequency and the threshold in quiet remained several decibels below the minimum of the pattern. Summarizing, the best range for test tones was; for the duration T_T of the test tone a range of 1/5 $T_M > T_T >$ 1/10 T_M, the repetition rate can be $f_P = f_M$ and for the test tone frequency f_T, $10f_M <$ $<f_T<20f_M$.

After establishing the best parameters for the test probe, masking results for a 50, 100 and 200 Hz tone were obtained. These results are shown in figure [2.3-46] in a comparative way that might be used to derive data for constructing a model of masking period patterns. The maximal level L_{Tmax} reached within the pattern [2.3-46a] as well as the temporal position Δtmax [2.3-46b] of this maximum within the period are plotted as a function of the masker tones. The parameter is the masker frequency f_M. As seen in figure 2.3-46a and for values about 10 dB above threshold in quiet, a square low relation exists between the level of the masker L_M and $L_T$max (dashed-dotted line). Although the 100 and 200 Hz maskers produced the same levels of threshold maxima, the 50 Hz tone produces maxima that are the same levels as the rest of the maskers only if the masker level is elevated by 8 dB.

From figure (2.3-46b) can be seen that the temporal location of the maxima is level dependent.

The dependence of the difference ΔL_T between the maximum and the minimum of the masking period pattern upon the masker frequency is shown in figure [2.3-47]. Note that parameters were not kept the same for all points, but rather, they were chosen differently for each case, so that they were optimal with respect to test tone attributes as was determined at the beginning of the experiments (i.e. conditions were sought which would produce maximal ΔL_T at each case). Whilst most subjects produced similar results one subject (dots) showed results which differ consistently showing that individual differences exist. From figure [2.3-47] it is clear that masking-period patterns show a similar shape for masker frequencies below 100 Hz, whilst for higher frequencies the difference between maxima ad minima becomes progressively smaller. At masker frequencies of about 500 Hz the masking period pattern for most subjects becomes flat, whilst a few subjects produce a pattern which only becomes flat at higher frequencies (1000 Hz). The masker frequency of 500 Hz has a period T_M = 2 msec. The strong decrement of ΔL_T with increasing masker frequency cannot be related to experimental conditions, rather there seems to be some kind of limit operating "psychoacoustically" in the ear.

In [37] more experiments were carried out with complex tones. These were either a distorted sine wave (obtained by adding to the fundamental a second and (or) third harmonic) or a beating tone by adding two adjacent higher harmonics of about the same level without any fundamental. In the latter case the phase relations play a secondary role, while they are most important in the former. In addition experiments using Gaussian-shaped impulses (having the property of small time and frequency spreading) as maskers were performed.

Comparisons were made between neurophysiological period histograms and masking period histograms. It should be noted that neurophysiological data are usually plotted on a linear scale whereas masking patterns are traditionally plotted on log scales. The two approaches are shown in figure [2.3-48]. These two figures also show the good agreement between the masking period patterns (m.p.p.) and the "rectified" time waveform.

Several mpp were therefore reploted on linear scales to allow a pictorial comparison with physiological period histograms, as well as with the time waveforms that were used as maskers. The masking effectiveness of the various components was used to derive appropriate scaling for each tone component before their time waveforms were added together to produce a comparison time waveform. In figure [2.3-49a,b] a 50 Hz sinosoid plus a 100 Hz sinosoid are scaled as above and added together to produce the time waveforms shown by dotted lines. The parameters for each graph are the level and phase of the second (100 Hz) harmonic when the first (50 Hz) was fixed at 100 dB SPL.

The agreement between the positive parts of the calculated time functions is very good, at least near the peak values. The masking period pattern and the calculated time function differ on the other hand, at such parts of the period where large negative values of the calculated waveform should lead to vary small values of the m.p.p. In contrast to this expectation, the patterns show a distinct rise in these areas, pointing to the fact that the condensation part of the time function also produces an increase in the threshold. However this increase is much less than that produced by a corresponding rarefaction. From these and similar plots it seems that smaller maxima of the time function tend to produce peak values in the patterns which are smaller than expected. This may be due to the fact that the relationship between mpp and the time waveform is a square law rather than linearity.

Finally in [38] an attempt was made to relate masking period patterns directly to the movement of the Basilar membrane rather than to physiological period histograms. In particular, the experiments aimed to determine whether m.p.p. are produced from p(t), the sound pressure at the eardrum or some derivative of this, $\dot{p}(t)$ (first derivative) or other e.g. $\ddot{p}(t)$. The first conclusion to the experiments was that at least for low frequency stimuli (<20 Hz) the whole of the BM is stimulated in phase. It was found necessary to produce a great many types of sounds with intricate waveforms which possessed some characteristics for P(t) (i.e. the actual time waveform) but not for $\dot{P}(t)$ or $\ddot{P}(t)$, or that $\dot{P}(t)$ was to produce a particular response that could not be accounted from P(t) or $\ddot{P}(t)$ etc: It was found that the second derivative $\ddot{P}(t)$ had direct correlation to the m.p.p., if it is assumed that positive values of $\ddot{P}(t)$ produce larger values of masked threshold than negative values of $\ddot{P}(t)$ of same size.

Some idealized patterns used and obtained in [38] are shown in figure [2.3-50]. These show that $\ddot{P}(t)$ is more relevant than P(t) or $\dot{P}(t)$ for the production of m.p.p. This was true for frequencies below 40 Hz but the m.p.p. were more correlated to the first derivative of pressure, $\dot{P}(t)$ for frequencies above 40 Hz. Also a careful study of the ensemble of mpp indicated that the higher peak in mpp belongs to a kind of suppression which would correspond to the displacement of the BM towards one direction, while the lower peak would belong to the excitation which would correspond to the displacement of the BM towards the other direction.

2.3.5 Discrimination

Discrimination describes the smallest perceivable difference between two stimuli, or alternatively, by how much a stimulus dimension can be changed before an observer can perceive a difference: The observer is presented with a stimulus having an attribute of magnitude S, where S may be its frequency or intensity or some other variable, and is asked to compare S with S + ΔS for a range of ΔS. For lower ΔS, S + ΔS is perceived as the same magnitude as S, but when a certain threshold ΔS is reached, S + ΔS is perceived as a new magnitude. This quantity is called the difference limen (DL) or the just noticeable difference (jnd). Note that the process here is nothing else than masking, since the stimulus with attribute S + ΔS can be split into the stimulus with an attribute S and a new stimulus of attribute ΔS. When S + ΔS is merely perceived as S we have complete masking of the stimulus (with attribute) ΔS by the stimulus S.

2.3.5.1 Intensity discrimination

The intensity DL is sometimes expressed in dB. This quantity is defined by:

$$\Delta I \text{ in dB} = 10\log_{10}[(I + \Delta I)/I] \qquad 2.3\text{-}12$$

which,for small ΔI becomes

$$\Delta I \text{ in dB} \approx 4.3\ \frac{\Delta I}{I} \quad \text{(through the power series expansion of the logarithm)} \qquad 2.3\text{-}13$$

An important concept in psychophysics is Weber's law which states that $\Delta I/I$ (the Weber fraction) is a constant (K) independent of I.

i.e. $\Delta I/I = K$: Weber fraction 2.3-14

This law is almost exactly true for a wideband noise, except near the absolute threshold of hearing. figure [2.3-51] [39]. It can be seen that for a range of nearly 90 dB, $\Delta I/I$ is constant.

For narrow band sounds the results are complicated. For sinosoids, Riesz [40], found that the Weber fraction $\Delta I/I$ changed little with the frequency of the sinosoid but it was not independent of intensity I i.e. $\Delta I/I$ was not a constant. Contrary to Weber's law $\Delta I/I$ decreased with increasing intensity, especially at low SLs and the curves became flatter at moderate and high SLs ($\Delta I/I$ approaching 0.3).

This discrepancy is referred to as "near miss" to Weber's law. More recent results [41], figure [2.3-52] confirmed Riesz's findings (at a frequency of 1000 Hz). In [42], through experiments on pulsed tones, it was found that the dependence of ΔI/I on frequency was even less than Riesz had measured and that a straight line could be used to show ΔI/I as a function of SL. (Straight line in figure 2.3-53). The existence of a DL for intensity stems from the probabilistic nature of the transformation from intensity to loudness. The path of the intensity information goes through the nerve fibers which show a probabilistic behaviour (neural spikes). In [a] p257 the magnitude of the DL is related to the inherent variability of the neural transduction.

2.3.5.2 Frequency discrimination

The most widely cited study of differential frequency sensitivity had been Shower and Biddulph [43], mainly due to the wide range of frequencies and sensation levels used in the study. A recent comprehensive study is that of Wier *et al.* [44]. Their results are shown in figure [2.3-54] together with some of those of the earlier study [43].

It can be seen that the DL for frequency Δf, is relatively flat for lower frequencies (<500 Hz) and this increases with frequency. Values as low as 1 Hz can be discriminated at most favourable ranges whereas at higher frequencies e.g 4000 Hz, 8000 Hz, it is roughly 16 Hz and 68 Hz respectively. Also note that the sensitivity changes with the sensation level (parameter in figure [2.3-54].

2.3.5.3 Discrimination of complex sounds

In [45] a study was undertaken to determine the number of frequency components that can be heard as "separate" from a multitone signal. Essentially, the subject was presented with a comparison single tone, and the multicomponent signal and asked whether the comparison single tone was also present in the multicomponent signal. (The exact procedure was slightly different to enable statistical analysis and numerical results). Plomp's results are shown in figure 2.3-55, together with the critical bandwidth as a function of frequency [10]. Results from harmonic complex tones (open points), inharmonic complex tones (solid points) and two tone signals (crosses) are shown. Note that the results are almost independent of the nature of the stimulus apart from the two tone signals at low frequencies, where the minimum frequency spacing to identify the two tones as being present, correctly, was two to three times smaller than with multicomponent signals. He concluded that the ear is able to distinguish a simple tone in a complex sound if the frequency distance to the adjacent tones exceeds the critical bandwidth. The same author also performed a masking experiment with a complex tone. The masker consisted of the first 12 harmonics of 500 Hz. The test tone was a 20 msec tone pulse. Its frequency was varied through all multiples of 50 Hz between 300 and 4000 Hz. The results are shown in figure 2.3-56. (it is worth noting that for each subject the measurements took about 10 hours!). It can be seen that the first five harmonics are "resolved" in the masking pattern, but not the higher harmonics. This agrees with the critical bandwidth being about 500 Hz around 3000 Hz.

More experiments with complex tones were performed in [46]. The experiments aimed to determine difference limens for frequency and intensity and were more in line with classical DL studies [42, 44] than Plomp's study [45] but were also concerned with complex sounds. The complex tones used in [46] all had fundamental frequencies of 200 Hz and all the harmonics were at 60 dB SPL. The different complexes used were formed from the following combinations of harmonics: 1-7, 1-12, 5-12, 6-12, 7-12. Only one harmonic component was changed at a time (either in frequency or intensity) and the DLs were determined in the presence of the other components. Therefore some masking from the rest of the complex sound was expected to be evident.

The results were as follows: The frequency DLs for the equal amplitude harmonics within the complex tone were small (less than 1%) for low harmonic numbers and increased markedly around the fifth to seventh harmonic. Nevertheless the highest harmonic in the complex was generally well discriminated.

The intensity DLs were again smallest for the lower harmonics and increased for harmonics above the fifth. The intensity DL for the highest harmonic was again small figure [2.3-57].

Finally to conclude the section on discrimination we report on some data on jnds for speech envelopes. Flanagan [47] undertook extensive investigations to integrate speech perception knowledge into formant vocoder design. Flanagan measured jnds for formant center frequency, bandwidth and intensity using sustained synthetic vowels. His results showed that the jnd for formant center frequencies was 3-5%, the jnd for bandwidth 20-40% and the jnd for intensity 1-3 dB. He also found that the jnd for intervalley intensity was around 10 dB and thus far larger than that for formant intensity. Recent studies [48] suggest that these values may be considerably higher for running speech. In [48] Klatt showed that for natural speech the jnd for fundamental frequency was 1.7% compared to that for stationary synthetic speech which was much lower 0.25%). Similar results could be expected for formant jnds. Also, the acuteness of the ear is known to decrease at higher frequencies for frequency discrimination with obvious implications for the higher formants.

A more recent study [49] assumed that all jnds are frequency dependent in the same fashion that frequency jnds are. They also assumed that the minimum jnds (at a frequency of 1.5 kHz) were multiples of Flanagan's values. The multiplication factor was found to be around 4 to 5. These results are summarized in figure 2.3-58. The jnds have the values 12% for formant frequencies, 80% for bandwidth, 4 dB for intensity and 10 dB for valley intensity, at a frequency of 1.5 kHz and follow the graph shown for other frequencies. Although this study showed that a distribution of jnds following a frequency dependent course gave better results than a flat distribution it was not examined whether one or more of the jnds could have had a flat characteristic with frequency whilst the others could have been frequency dependent. Their results were derived through a scalar LPC parameters quantization paradigm which also showed that the quantization based on jnds was superior to the one based on Log Area Ratios [50].The quality judgements were performed on the unquantized LPC residual filtered through the quantized filters. This method ignores the effect of quantization on the magnitude of the prediction error power, or other prediction error characteristics, which may be of equal if not greater importance in a real coding situation.

2.3.6 Phase perception

The first systematic study on phase perception was the classical work of Mathes and Miller [51]. The experiments centered upon small groups of harmonics of fundamental pitches in the voice range (But identical results were obtained with inharmonically related but equally spaced frequencies).

The effects observed were concerned with a change from rough to a smooth sounding quality as the envelope shape of the complex wave changed from one in which large maxima and minima existed in the time structure to one for which the envelope was more nearly uniform in time. It was found that the change from roughness to smoothness could be accomplished by changing either the amplitude or the phase of a particular component or set of components. Amplitude and frequency modulated signals were mostly used, but also other more complex envelopes for the sound were employed in the experiments.

They found that the degree of harshness was related to the relative length and depth of the recurrent depressions in the envelope wave. With the amplitude modulated wave, and stating from a rough signal, the roughness could be made to disappear by changes in the phase alone of one of the components of the complex wave. This change towards smoothness was also accompanied by a change to the envelope of the wave towards uniformity with time. A most important result was that these changes from rough to smooth occur only if the modulating frequency is lower than a fixed percentage of the carrier frequency (~ 40%). In order for the phase effect to be isolated it was necessary to use head phones for listening. With loudspeakers it was possible by changing the position of the head in relation to the standing wave patterns, to observe either the rough or smooth sensation. This may be one of the reasons why encoded speech segments sound different when heard through head phones rather than loudspeakers. Also, the fact that the separation of frequency components reduces roughness, can explain why female voice is generally regarded as more melodious than that of the male. This is so because in the former case a specific speech resonance (formant) includes fewer and lower order harmonics.

The next major study was performed by Zwicker but unfortunately the references are in German. Some of the results can be found in Goldstein [52] who confirmed and extended these results. Zwicker performed a variant of the AM/FM phase-perception experiment in which he found that a carrier tone with just noticeable sine amplitude modulation has essentially the identical spectral composition as a similar carrier with just noticeable sine frequency modulation, provided that the modulation frequency exceed some critical value. The critical value is dependent only upon the carrier frequency and not upon its amplitude. It was found to be about half the critical bandwidth [12]. Since phase is the only significant difference between the AM and particular FM stimuli used, that could cause differences in the threshold perception of the different modulations, one concludes that phase is ignored when the modulation frequency exceeds half a critical band or, equivalently, when the stimulus bandwidth exceeds one critical band. Hence limited auditory frequency resolution seems to be responsible for phase perception. Some of Zwicker's results are shown in figure [2.3-59]. In [52] Goldstein reports that Mathes and Miller experiments [51] indicated a critical modulation frequency of 40% whereas Zwicker found a critical value of 10% This was attributed to the differences in the experiments which indicated that for these effects auditory filtering may not be modelled in general by filtering with an ideal rectangular passband filter with critical bandwidth. Goldstein reported the two sets of experiments from the two sources.

For the Mathes and Miller experiment the stimuli could be described by the following equation

$$\frac{1}{2}\cos\ (w_1-w_2)\ t + \cos\ (w_1 t + \phi) + \frac{1}{2}\cos\ (w_1+w_2)t \qquad 2.3\text{-}15$$

where the phase angle ø is zero for AM and Π/2 for QFM (quasi - FM). The carrier frequency is w_1 the modulation frequency is w_2.

The two stimuli are shown in figure 2.3-60. The modulation frequency was varied for a range of SPLs and subjects were asked to decide whether the perceived sounds were identical or different. The results are shown in figure [2.3-61]. The critical band scale [12] is also shown. The course of the data versus carrier frequency is remarkably similar to the critical-band function but a level dependence is also evident. For modulation frequencies in the lower region of the graph, the QFM was perceived as temporally much smoother and steadier than the AM. Below modulation frequencies of about 20 Hz a warble or pitch fluctuation could be heard in the signal. The AM signal was described as having loudness fluctuation, a chirp, or a buzz. This holds up to a frequency of 4 KHz. As the carrier is increased beyond this frequency the quality distinctions become increasingly more subtle.

Subsequently, experiments similar to those performed by Zwicker were carried out. An AM signal with modulation depth m can be expressed as:

$$\frac{m}{2} \cos (w_1-w_2)t + \cos w_1 t + \frac{m}{2} \cos (w_1+w_2)t \text{ : AM} \qquad 2.3\text{-}16$$

A QFM signal will modulation depth b may be defined as:

$$\frac{b}{2} \cos (w_1-w_2)t + \cos(w_1 t - \frac{\Pi}{2}) + \frac{b}{2} \cos (w_1+w_2)t\text{: QFM} \qquad 2.3\text{-}17$$

which approximates a sine FM tone with small modulation index (b^2<<1).

$$\sin (w_1 t + b\cos w_2 t) \text{ :FM} \qquad 2.3\text{-}18$$

The threshold FM signals in Zwicker's results figure [2.3-59] are approximated by QFM at most modulation frequencies. Therefore any differences in the sensation of the two stimuli must be attributed to the phase angles of the stimuli. Zwicker's experiments were simulated using QFM instead of FM. The results were very similar to those obtained in Zwicker's study, figure [2.3-59].

Goldstein concluded that the results could be explained if the ear was operating as in practical spectrum analysers [47] where the short time amplitude spectrum is obtained through a set of bandpass filters followed by quadrature detection which is approximated by ideal envelope detection. Further, he showed that the modulation is optimally detected on the steepest part of the bandpass filter and that the slope of this part of the filter was the only relevant feature of the filter related to modulation thresholds (and hence phase detection). The relation between modulation thresholds and critical bands simply means that this filter slope is (approximately) independent of place (frequency) when frequency is expressed in critical band units or any other unit proportional to the critical band scale. The value of this slope was about 30 dB per critical band.

The ideal envelope detection hypothesis predicts that "linear" phase transformations of the type:

$$\theta' = \theta + \alpha + bf \qquad 2.3\text{-}19$$

where θ' and θ are the new and old angles respectively, f is the frequency and α, b, are arbitrary constants, would produce identical short-time spectra [52,53]. The term bf is trivial since it only causes a simple delay to the whole spectrum. The term α though will alter the waveform although leaving the envelope intact. Although some experiments in monaural phase perception support the envelope detection hypothesis others do not. The divergence of the ear's behaviour from this model could perhaps be accounted for by known nonlinearities in the ear (combination tones) [53]. Finally, it should be noted that the experiments carried out in [35-38] concerning masking period patterns are also very relevant to, and are examples of, phase perception.

2.3.7 Roughness

We have seen that phase differences become detectable through changes in the roughness of sounds. Roughness is a distinct sensation or attribute of a sound as is its loudness or pitch: If a steady sound is amplitude modulated with a low frequency of e.g. 5 Hz, one easily recognises the amplitude fluctuations as corresponding fluctutations of lougness. As the modulation frequency is increased beyond about 20 Hz, the fluctuations are still well perceived, however one does no longer distinguish the succeeding maxima and minima as separate events. Rather, the loudness of the sound is constant, and the fluctuations are perceived as an unpleasant disturbing component which usually is called "roughness", "raucousness" or "harshness". Studies on roughness attempt to measure the magnitude of roughness and its dependence upon the stimulus's physical charcteristics. One such typical study is from Terhardt [54]. It is known that for AM signals the magnitude of roughness depends upon the modulation index. To determine a quantitative relationship Terhardt performed an experiment where the subjects had to listen to two AM tones which differed by their modulation indices (m). In one experiment the subjects had to judge whether the second AM tone was more or less than half as rough than the first one. From the distributions of answers, those m-values of the second AM tones were determined at which the answer "more than...." and "less than...." had equal probabilities. Those m-values were considered as the resulting values $m_{0.5}$

corresponding to half the roughness of an AM tone with the m value m_1. In another experiment subjects were asked to determine the tone being "twice as rough" as the test tone. This was now the m_1 tone where the test tone was $m_{0.5}$. Three AM tones were investigated (1) carrier frequency f = 268 Hz, modulation frequency f_{mod} = 40 Hz; (2) f = 1 KHz, f_{mod} = 70 Hz; (3) f = 4 KHz f_{mod} = 70 Hz. The SPL was 70 dB in all cases. The results are shown in figure [2.3-62]. The results reveal that $m_{0.5}/m_1$ does not depend on m_1, carrier frequency or whether is determined with the criterion "twice as rough" or "half as rough".

The arithmetic mean of all resulting values $m_{0.5}/m_1$ shown in the figure is 0.707. This means that on the average, the degree of modulation has to be changed by a factor of 0.7 in order to change the roughness by a factor of 1/2. Therefore the relation between roughness r of a sinosoidally amplitude modulated tone and the degree of modulation m can be described by the equation

$$r = \text{const. } m^2 \qquad 2.3\text{-}20$$

The SPL of the AM tone has an effect upon roughness, but an SPL variation of 40 dB (with the m value being constant) produces a smaller difference as the change of the m value by a factor of two. The exact relationship depends heavily upon the sequence of the presentation of the test tones.

Comparisons between FM and AM tones yielded similar results to that of [52]. The "const" in the equation mentioned above depends on center and modulation frequency. The relative roughness as a function of modulation frequency f_{mod}, with carrier frequency, f_{car} as parameter is shown in (fig. 2.3-63) for a given SPL and modulation index. At low modulation frequency no sensation of roughness is created although a loudness fluctuation can be perceived. Roughness seems to begin at about 20 Hz and rises very quickly for higher modulation frequencies. The peak value is reached at modulation frequencies which depend on the carrier frequency. For carrier frequencies below 2 KHz the maximum is reached at lower modulation frequencies and is not as high as for carrier frequencies above 2 KHz. For the higher frequencies, the modulation frequency for maximal roughness as well as for vanishing roughness does not depend on the carrier frequency. For these frequencies the roughness peak is reached at 70-80 Hz modulation. For higher modulation frequency the roughness decreases quickly and reaches, at 250 Hz, 1/10 of the peak value which means that it almost vanishes.

The jnd of roughness as a function of modulation frequency for an AM tone can be seen in figure 2.3-64 for a 4 KHz center frequency. The dashed line shows the modulation jnd. The two are different over the higher modulation frequency range since effects other than roughness serve to perceive modulation at these frequencies. The monotonic increase of the roughness threshold with growing f_{mod} must be ascribed mainly to the influence of "low pass characteristics" of the neural system since the AM tone's amplitude spectrum is well within the critical bandwidth (≃ 700 Hz). The neural system seems to transmit the envelope fluctuations of the neural stimulus with a frequency characteristic similar to that of a simple RC-network with a time constant of about 13 ms (solid line in figure 2.3-64). From this figure the maximum modulation frequency that can be heard is around 300 Hz. From experiments on more complex sounds he deduced that for the summation of roughness of complex sounds the following are true:

The entire roughness is composed of the partial roughnesses which are contributed by adjacent critical bands. The entire roughness is therefore the sum of the partial roughnesses which are determined by taking into consideration the effective degree of modulation in each band, after the effects of the critical bandwidth and the low pass characteristic (figure [2.3-64]) are taken into account.

In [55] Vogel through masking experiments reached the conclusion that the fluctuation of the entire neural activity pattern (excitation pattern) was responsible for the sensation of roughness and that roughness over a large frequency range can be summed in the same way as loudness to give an overall sensation as will be seen in the section on psychoacoustic models.

2.3.8 Pitch

Like loudness or roughness the word pitch denotes a perception with which we are all familiar. Pitch is the psychophysical correlate of frequency such that high frequency tones are heard as being "high"in pitch and low frequencies are associated with low pitches. A scale for pitch can be obtained in a similar manner as for other psychophysical attributes, by requesting subjects to find stimuli with a pitch half as that or twice as that of a standard. Several other methods can be employed to obtain a pitch scale by using different relationships between stimuli and requesting different tasks to be performed by the subjects. Note that different methods yield slightly different results. The unit of pitch is the mel. By convention, 1000 mels is the pitch of a 1000 Hz tone presented at 40 phons. The intensity is specified since pitch also depends slightly upon stimulus intensity. One such scale relating pitch to frequency is shown in figure 2.3-65. which is taken from [56]. Approximately 150 mels correspond to the critical bandwidth [57].

A single tone gives rise to a pitch sensation which is usually refered to as spectral pitch and is strongly correlated to its frequency. A complex sound on the other hand (e.g. a selection of harmonics) also gives rise to a pitch which can be matched to the pitch of a pure tone, but the relationship between the psychophysical sensation and the physical components of the stimulus is less clear. It is possible for example for a complex sound to have a pitch matched to a sinosoid of frequency F without the complex sound itself having a component near or at F. This pitch phenomenon is refered to as virtual pitch, periodicity pitch, or residue.

The mechanisms by which the physical stimulus is processed by the auditory system to produce the sensation of residue have been the argument of heated debates for over a century now and various theories have been proposed. It seems that after the preprocessing stage whose effects are common to all other psychophysical phenomena, a pattern matching procedure is followed to determine virtual pitch. An excellent review of the subject is given in [58].

Although a study of the mechanisms of virtual pitch is beyond the scope of this work we will report on a few experiments that may be of interest to speech which, after all, is an example of the virtual pitch phenomenon and speech coding, especially to a class of coders employing regeneration of the harmonic structure in certain frequency bands at the decoder, by means of spectral folding or spectral translation [59].

We have seen that the ear performs some kind of bandpass frequency analysis and that the waveforms within these bands are to a certain extent preserved. The outputs from one such bank of bandpass filters is shown scematically in figure 2.3-66. The input is a periodic pulse train of period 200 Hz. It can be seen that the lower channels resolve the individual harmonics, whereas the upper channels reflect the periodicity of the input waveform. Assume now that individual harmonics that fall into one of the higher frequency filters are shifted in frequency by a constant amount i.e.

$$f'_i = f_i + \Delta f = i.f_o + \Delta f \qquad 2.3\text{-}21$$

where f'_i is the new ith harmonic, f_i the old ith harmonic, f_o the funadamental and Δf the frequency shift. It is easy to deduce that the output waveforms shown in figure 2.3-66 will change very little. This follows from the fact that the same pattern is now "seen" by a filter centered at $F + \Delta f$ which was "seen" by a filter centered at a frequency F before the frequency shift.

Since the outputs of successive filters at higher frequencies are more or less similar (with respect to envelope periodicity) the deduction follows. The frequency shift by Δf for a number of harmonics corresponds exactly to the regeneration process mentioned above [59].

De Boer [60] performed some experiments concerning such "frequency shifted" harmonic signals. His signals only involved the higher harmonics and were generated by a modulation process and carrier injection similar to AM. For a carrier frequency f and modulation frequencies g, 2g and 3g, the following components were obtained

f-3g, f-2g, f-g, f, f+g, f+2g, f+3g

g was held constant at 200 Hz throughout the experiment and f was varied. At f = 2000 Hz a harmonic series is obtained

1400, 1600, 1800, 2000, 2200, 2400, 2600 Hz

This is because f = 2000 Hz an integral multiple of g = 200 Hz. When f = 2200 Hz then, again, a harmonic series is obtained. For intermediate values of f the situation of "frequency shifted" harmonics or inharmonic spectra is obtained. De Boer notes that the inharmonicity was not observed at all; the residue did not sound much different, but the pitch was slightly different. For example at f = 2030 Hz the components are at

1430, 1630, 1830, 2030, 2230, 2430, 2630 Hz

The residue is found to be tonal for this combination, with a slightly different pitch of 205 Hz. Figure [2.3-67] shows the results of pitch matches as a function of the ratio f/g. There are two pitch courses visible in the figure, and, somewhere in the centre of the interval, the attention of the listener seems to switch over from one to the other pitch. Clearly pitch is ambiguous in this region. Both of these pitches however are near 200 Hz. For the middle point f = 2100 Hz, the components are

1500, 1700, 1900, 2100, 2300, 2500, 2700 Hz

This waveform has a tone periodicity of 100 Hz. The pitch that is heard though is near 200 Hz as for the rest of the combinations.

The waveforms of the three conditions f = 2000 Hz, f = 2030 Hz, f = 2100 Hz are shown in figure [2.3-68]. It can be seen that the envelopes are the same (same modulation components at g, 2g, 3g) but that a pseudo period near to the envelope period can be seen corresponding to the "sampling" of the envelope by the fine structured waveform. This could be a clue as to how the ear perceives the pitch of such complex tones.

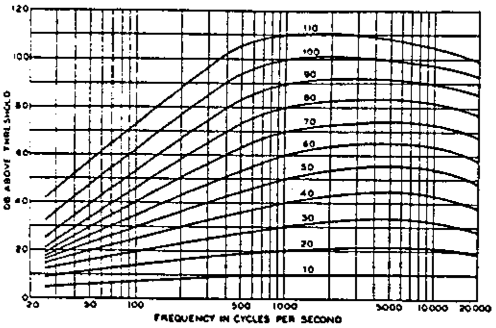
In the above experiments only the higher harmonics were present. It is important to know what happens when the low frequency harmonics are also present but not shifted. Although this exact condition was not examined a similar condition where a complex sound consists of lower harmonics belonging to a fundamental frequency f_o up to a cutoff frequency f_c and then upper harmonics belonging to a different fundamental $f_o + \Delta f$ was set up in [61].

The stimulus structure is shown in figure 2.3-69. The experiments found that the pitch corresponded to the part of the sound with the harmonics at the lower part of the spectrum. They found that even a foursome of harmonics of frequency f_o against a multitude of components (all components of a pulse train above the fourth) with a fundamental $f_o + \Delta f$ forced the total sound to have a pitch corresponding to f_o. The lower components assumed dominance once they were above a sensation level of 10 dB SL.

The above two experiments indicate that frequency shifts of harmonics in the upper bands of the speech spectrum would have little perceptual impact as indeed is the case as found in [59]. (See also chapter six).

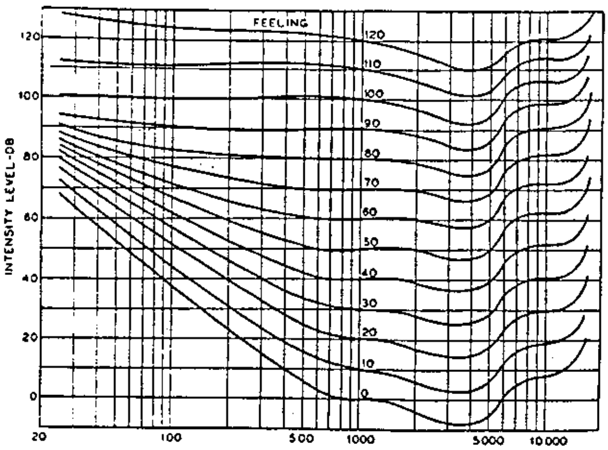
2.3.9 Timbre

Timbre is the attribute of auditory sensation in terms of which a listener can judge that two sounds having the same loudness and pitch are dissimilar [62, 63]. Timbre is that sensation which distinguishes between two different instruments e.g. the violin and the piano when playing the same note at the same loudness. The most important verbal attribute of timbre is sharpness, [64] which in turn is largely determined from the characteristics of the spectral envelope. It is thus important for the classification of speech sounds. [65, 66, 67]. It is difficult to define exactly what timbre is, and although the short time envelope is one of the strongest physical correlates, phase information (i.e. the time course) of the stimulus is also relevant [62]. The difficulty to define timbre is perhaps understandable since it encompasses any hearing sensation that is not "loudness" or "pitch". It is therefore expected to have many physical correlates. This is reflected in the fact that multi-dimensional analysis techniques are often employed in the study of Timbre.



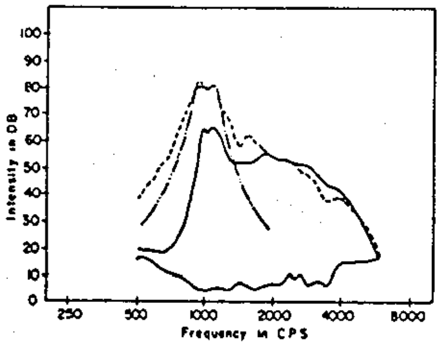
—Loudness Level Contours versus Sensation Level.

Figure 2.3-1 [b]



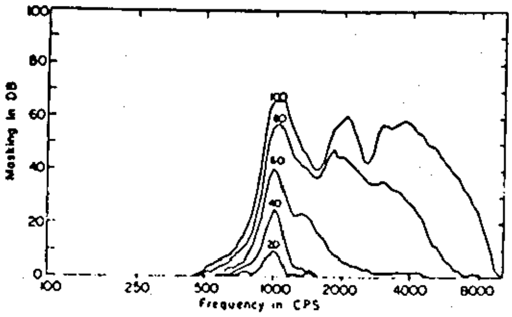
Loudness Level Contours versus Intensity Level.

Figure 2.3-2 [b]



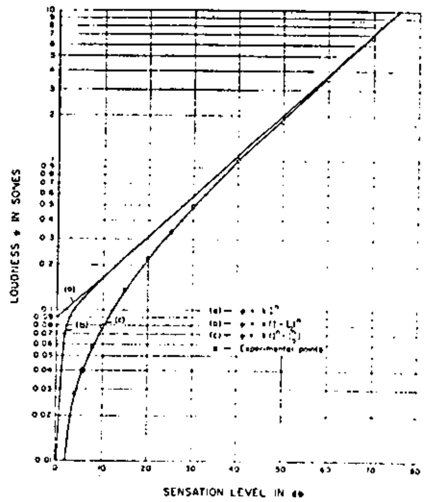
Masked thresholds for 1000 cps, 80 db SL. The over-all SPL of the tone and noise band are 87 db. The frequency limits of the noise band are 880 and 1150 cps and the SPL per cycle is 63 db.

Figure 2.3-3 [9]



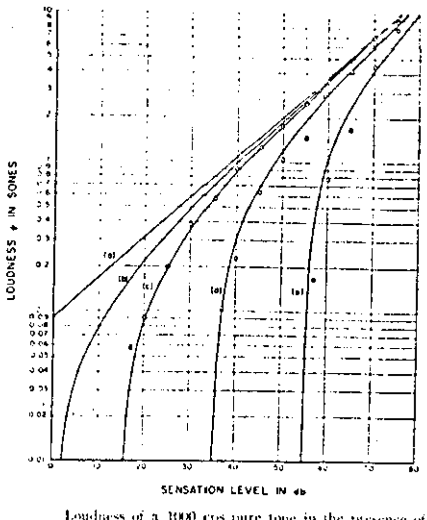
Average masking patterns for 1000 cps based upon three listeners. The sensation level of the masking tone is attached to each curve.

Figure 2.3-4 [6]



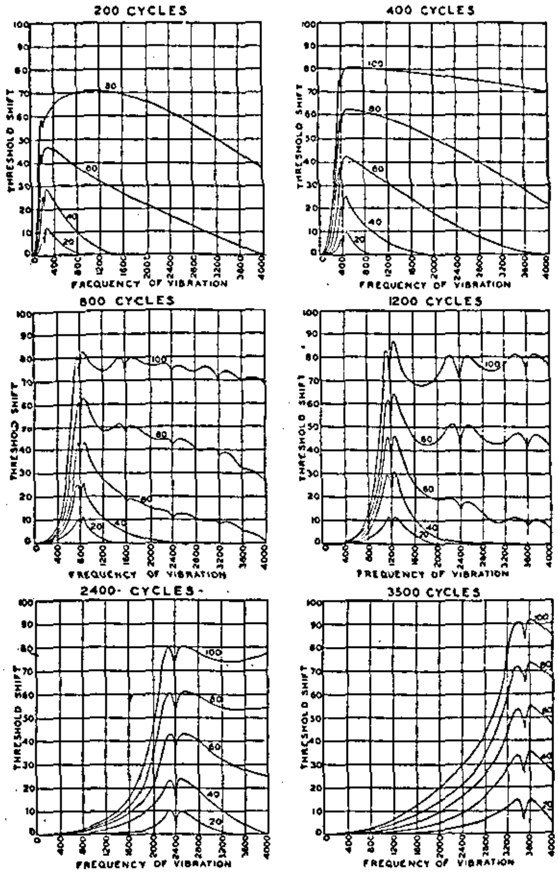
Loudness curves given by the relations $\psi = kI^n$, $\psi = k(I - I_0)^n$, and $\psi = k(I^n - I_0^n)$. The experimental points were obtained by Hellman and Zwislocki.

Figure 2.3-5 [3]



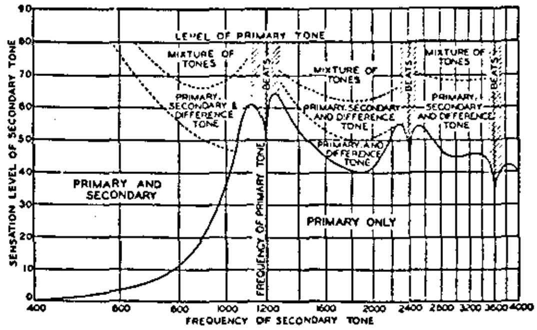
Loudness of a 1000 cps pure tone in the presence of physiological noise [curve (b)] and octave band random noise that gives pure tone threshold levels of 15(c), 35(d), and 55 db(e). Experimental points are shown on the latter three curves obtained from the relation $\psi = k(I^n - I_0^n)$. The straight line (a) is the curve for $\psi = kI^n$.

Figure 2.3-6 [3]



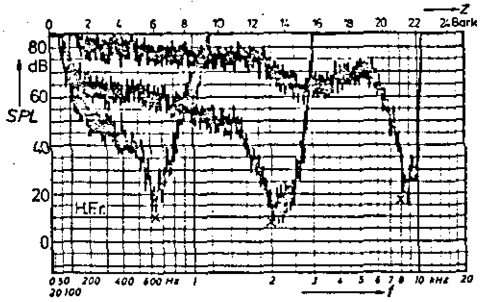
MASKING DATA FOR 200, 400, 800, 1200, 2400, AND 3500 CPS.

Figure 2.3-7 [b]



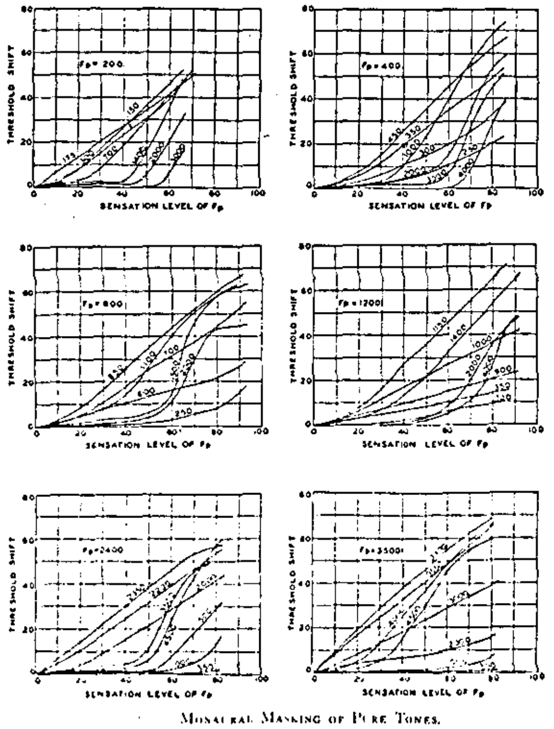
SENSATION LEVELS CAUSED BY TWO PURE TONES.

Figure 2.3-8 [b]



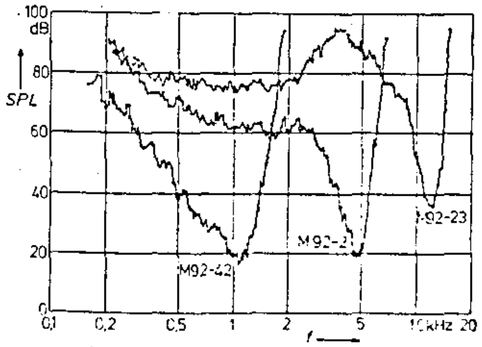
Original recordings of psychoacoustical tuning curves(SL=5dB) for characteristic frequencies of 630 Hz, 2kHz and 8kHz. abscissa: frequency f (lower scale) and critical band rate (upper scale), respectively.

Figure 2.3-10 [7]



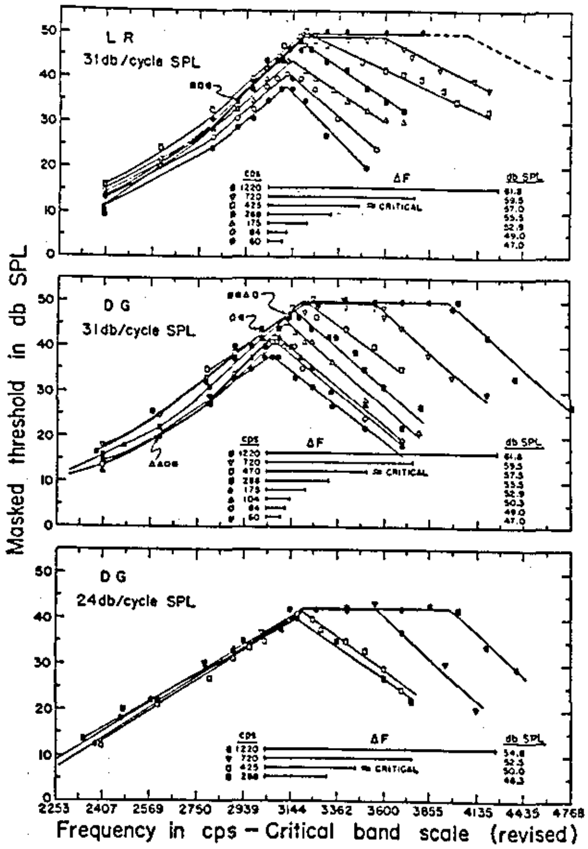
MONAURAL MASKING OF PURE TONES.

Figure 2.3-9 [b]



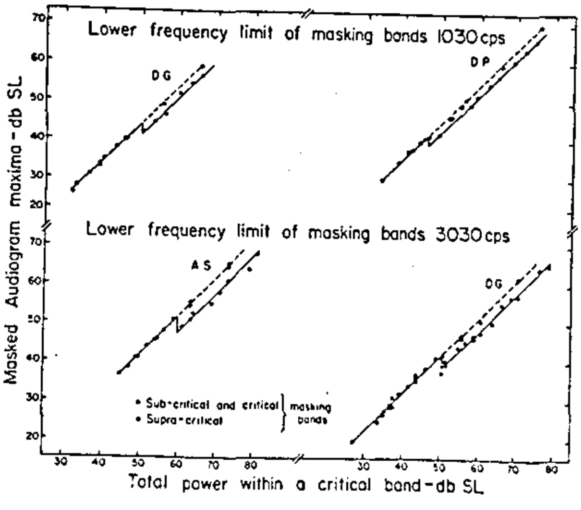
Computer processed tuning curves from single auditory nerve fitres (cat) of different characteristic frequencies

Figure 2.3-11 [7]



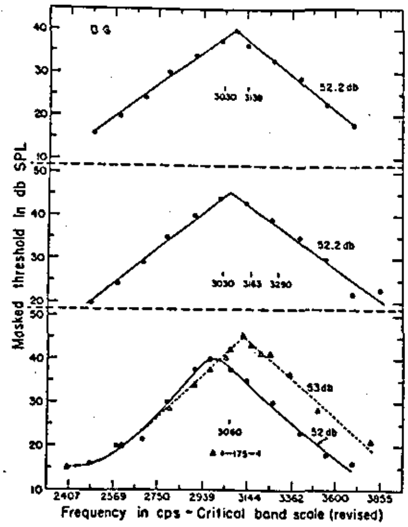
Masked audiograms as a function of increasing bandwidth for subjects LR and DG. The spectrum level and the lower frequency limit of the masking noise were held constant. The spectrum level appears at the upper left of each box. The lower limit of all bands was 3030 cps. Bandwidths are in cps and in the form of horizontal bars showing the position of the bands on the frequency scale. Four abscissa intervals represent a critical band. The level of each band of noise is given at the end of each bar. In these and subsequent figures, SPL = spectrum level +10 log(ΔF-20), where 20 cps is the width of the hole in the center of each band of noise. The heights of the audiograms are approximately proportional to the total power in the masking bands for bandwidths up to critical width.

Figure 2.3-12 [11]



The points on the graph are the heights (peak thresholds) of masked audiograms plotted as a function of level. Thresholds are expressed in decibels above absolute threshold, or amount of masking. The solid-line two-segment curves show the masking produced by subcritical and critical bands as a function of the sensation level of the bands. The dotted lines are the upper continuations of one-segment curves that plot the maximum threshold of the flat-topped masked audiograms produced by supracritical bands of noise. Masking is plotted against the sensation level of a critical fraction of the total bandwidth of the supracritical stimuli. The one-segment curves coincide with the lower segments of the two-segment curves but continue upward to the right when the transition level is exceeded. At 3030 cps the ordinate intercepts for the two subjects are almost the same; at 1030 cps the intercepts for the two subjects differ by 1.0 db.

Figure 2.3-13 [11]



Masked audiograms produced by line spectra. The dotted audiogram produced by a 175-cps band of noise is included for comparison. To the right of each audiogram is given the total power of the masking stimulus. The frequencies of the components of the masking stimuli are given below each audiogram. The masked signal was a 60-cps band of noise, approximately equal to one-seventh of a critical band. Signal threshold was plotted as a function of center frequency. The single and two-tone audiograms are nearly the same height but the three-tone audiogram is the same in height as the noise audiogram.

Figure 2.3–14 [11]

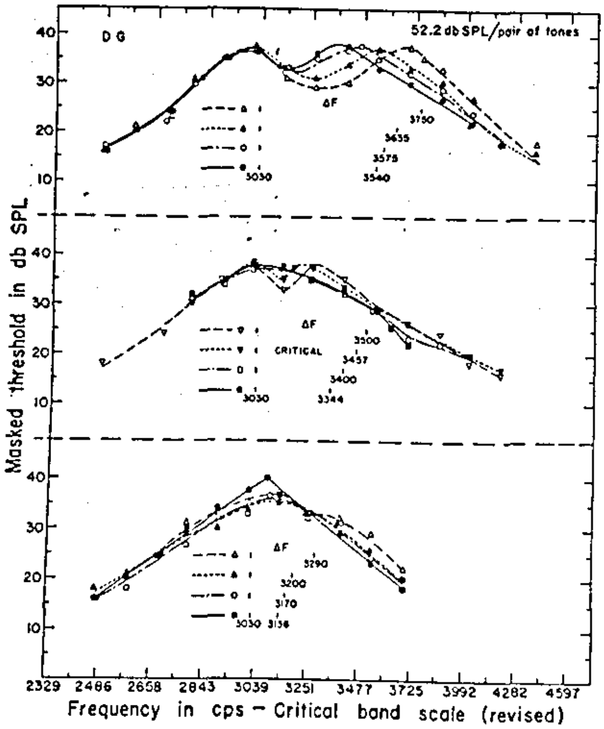


Figure 2.3-15 [11]

Twelve masked audiograms produced by a two-tone stimulus as a function of the frequency separation of the two tones. Component frequencies of the masking stimuii are given below each set of audiograms, which are shifted on the ordinate for clarity of presentation. The masked signal was a 60-cps band of noise, approximately equal to one-seventh of a critical band. Signal threshold was plotted as a function of center frequency. The dip in the audiogram appears below the midpoint of the interval separating the tones when the separation of the tones reaches or just exceeds critical width. As the separation grows, the two peaks come to resemble closely the peak of the audiogram produced by a single pure tone.

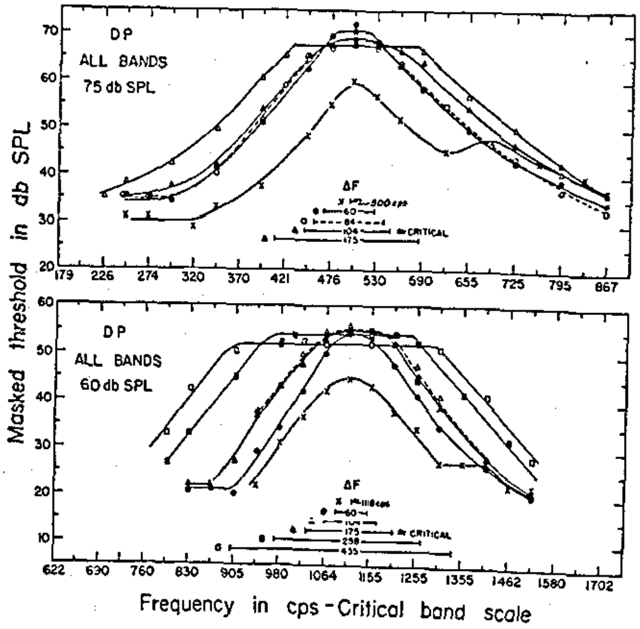
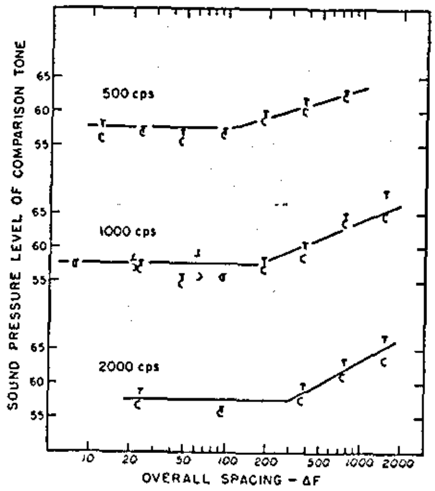


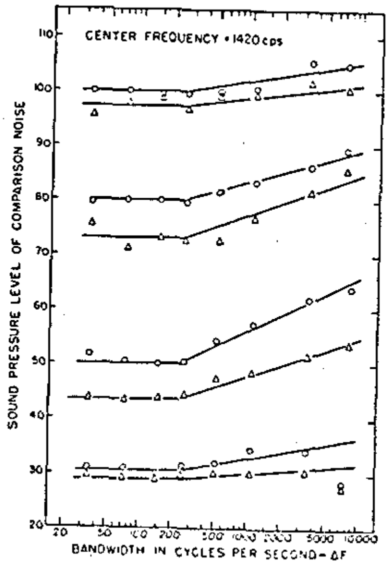
Figure 2.3-16 [11]

Masked audiograms of subject DP produced by bands of noise of equal power and the same center frequency. The bands were centered around 500 and 1118 *cps*; bandwidths ranged from subcritical to supracritical width. To determine the masked audiograms produced by pure tones 500 and 1118 cps, the signal used was the 60-cps band of noise shown at the bottom of the graph; the threshold of the noise was plotted as a function of the center *frequency*. Bands of critical or lesser width produced audiograms that are nearly superimposable, tending to coincide most closely near the abscissa; those produced by pure tones of equal power are quite comparable in extent. The extra extent of audiograms produced by supracritical bands is about equal to the amount by which the bands exceed critical width.



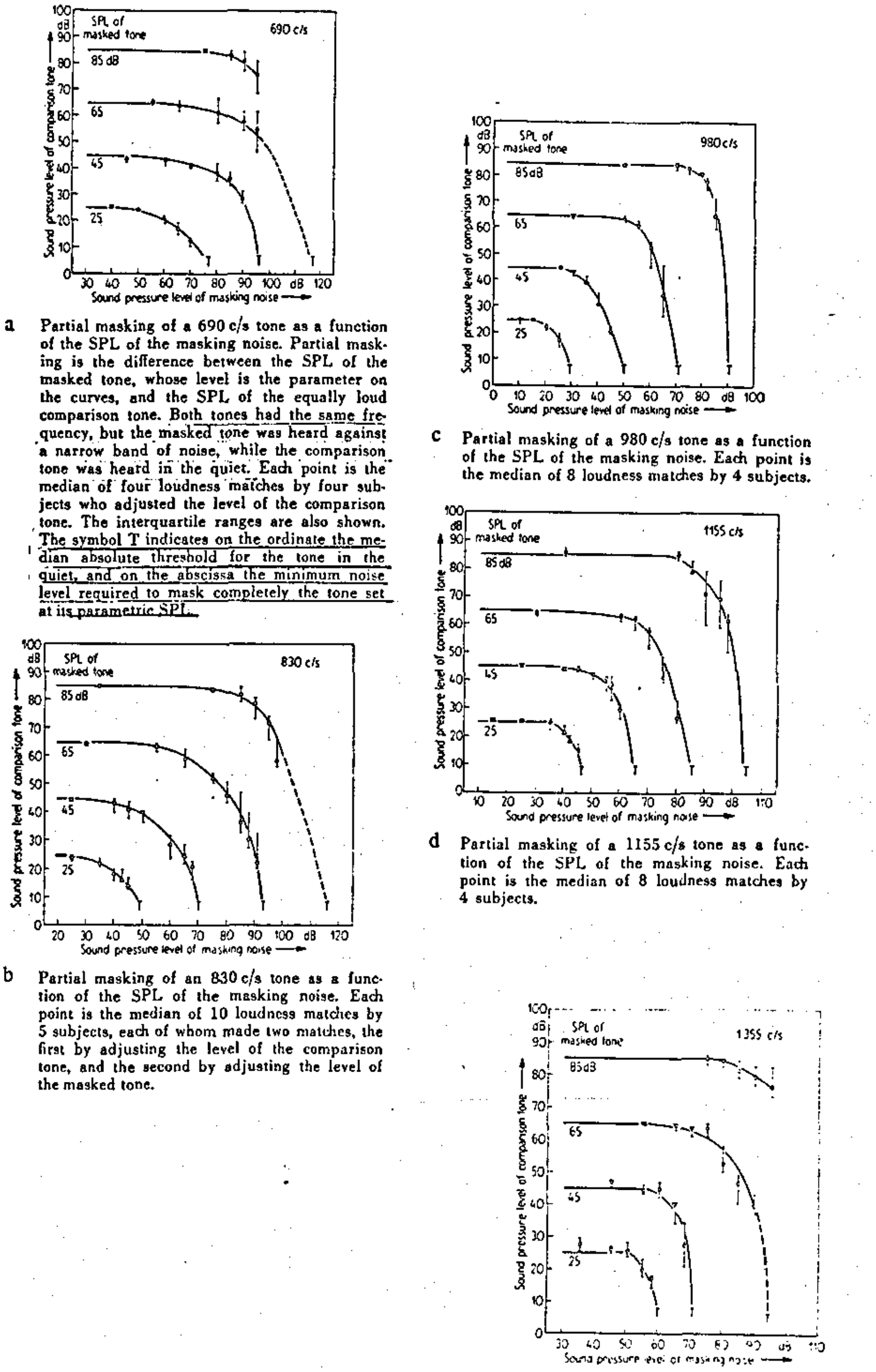
Dependence of loudness on the spacing of the components in a four-tone complex. *The four tones were spaced approximately* uniformly in frequency about the center frequency indicated. Loudness balances between the center frequency and the complex were made by groups of from 16 to 22 subjects. *T* means the tone was adjusted and *C* means the complex was adjusted. The lines through the data have a break at the point predicted by the critical-band hypothesis.

Figure 2.3-17 [10]



Dependence of loudness on the band width of a noise of constant SPL having a center frequency of 1420 cps. Comparison noises of two band widths, 210 cps (circles) and 2300 cps (triangles), were matched in loudness to each band width *ΔF* at a constant SPL.

Figure 2.3-18 [10]



a Partial masking of a 690 c/s tone as a function of the SPL of the masking noise. Partial masking is the difference between the SPL of the masked tone, whose level is the parameter on the curves, and the SPL of the equally loud comparison tone. Both tones had the same frequency, but the masked tone was heard against a narrow band of noise, while the comparison tone was heard in the quiet. Each point is the median of four loudness matches by four subjects who adjusted the level of the comparison tone. The interquartile ranges are also shown. The symbol T indicates on the ordinate the median absolute threshold for the tone in the quiet, and on the abscissa the minimum noise level required to mask completely the tone set at its parametric SPL.

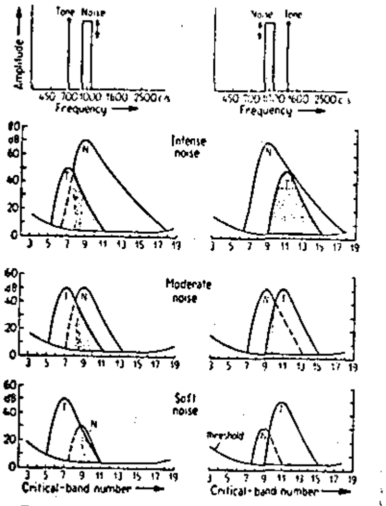
b Partial masking of an 830 c/s tone as a function of the SPL of the masking noise. Each point is the median of 10 loudness matches by 5 subjects, each of whom made two matches, the first by adjusting the level of the comparison tone, and the second by adjusting the level of the masked tone.

c Partial masking of a 980 c/s tone as a function of the SPL of the masking noise. Each point is the median of 8 loudness matches by 4 subjects.

d Partial masking of a 1155 c/s tone as a function of the SPL of the masking noise. Each point is the median of 8 loudness matches by 4 subjects.

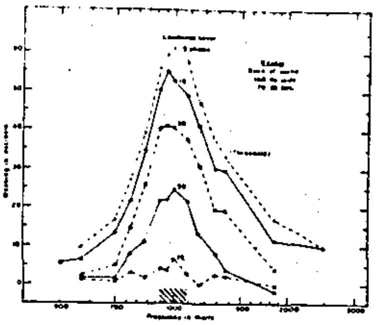
c Partial masking of a 1355 c/s tone as a function of the SPL of the masking noise. Each point, except on the 65 dB curve, is the median of 8 loudness matches by 4 subjects. On the 65 dB curve, each point is the median of 6 matches by 3 subjects. The fourth subject had an unusually high masked threshold for the 65 dB tone so that his loudness judgments were considerably below those of the other three subjects and were not included in the results.

Figure 2.3-19 [15]

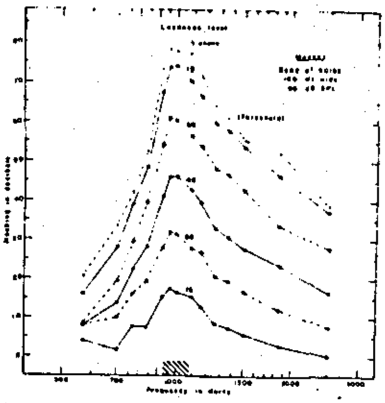


Excitation patterns for the masking noise and two of the tones used in the present experiments. The idealized spectra for the noise and tones are also shown. As the intensity of the noise decreases, those overlapping areas (shaded) in which the excitation level of the noise is greater than that of the tone also decreased in size. The decrease is greater for the tone lying above the frequency range of the noise than for the one lying below, which agrees with the finding that partial masking changes more rapidly for higher than for lower frequency tones.

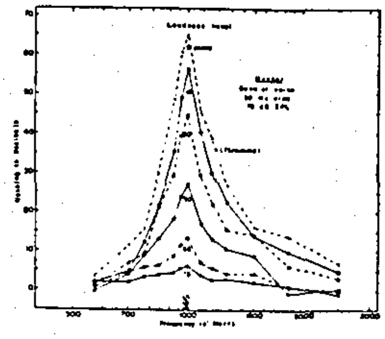
Figure 2.3-20 [15]



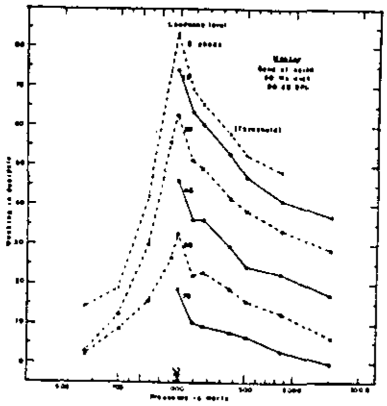
a Complete and partial masking by a band of noise, 160 Hz wide at 70 dB SPL.



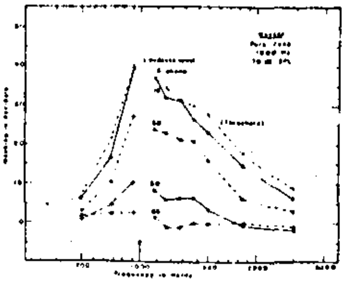
b Masking by a band of noise, 160 Hz wide at 90 dB SPL.



c Masking by a band of noise, 50 Hz wide at 70 dB SPL.



d Masking by a band of noise, 50 Hz wide at 90 dB SPL.



C Masking by a 1000-Hz tone at 70 dB SPL.

Figure 2.3-21 [16]

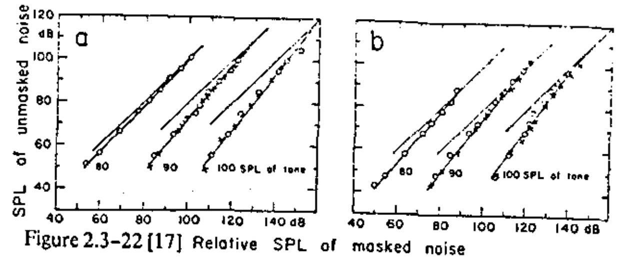
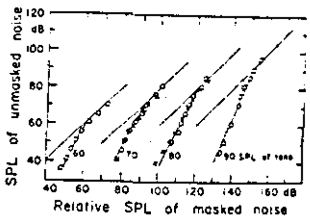


Figure 2.3-22 [17] Relative SPL of masked noise

(a) Loudness-matching functions obtained by equating in loudness a masked and unmasked band of noise. Noise bandwidth: 75-9,600 Hz. Masking frequency: 1,000 Hz. The crosses indicate group geometric means obtained by adjustment of the unmasked noise, and the unfilled circles indicate those obtained by adjustment of the masked noise. The parameter is the SPL of the 1,000-Hz masking tone. The numbers along the abscissa indicate the SPLs of the noise masked by the lowest intensity tone. To determine the SPLs of each successive function, subtract 20 or 40 dB, respectively, from the abscissa values (b) Analogous to (a), except that the noise band was 600-1,200 Hz wide.



Analogous to Fig. 1a, except that the noise band was 925-1,080 Hz wide. To determine the SPLs of the noise masked by the 90-dB SPL tone, *subtract 60 dB* from the abscissa values.

Figure 2.3-23 [17]

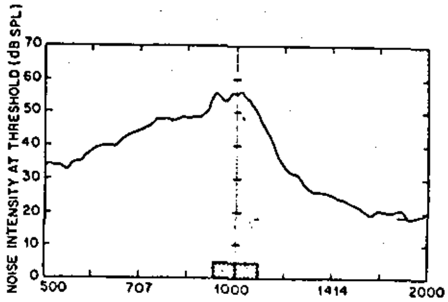


Figure 2.3-24 [19] FREQUENCY OF MASKING TONE (HZ)

Auditory threshold for a critical-band noise burst centred at 1 kHz masked by a tone of intensity 80 dB SPL. The frequency band occupied by the noise is indicated by the rectangular shaded area. Note that for a tone frequency of 1 kHz the noise intensity at threshold is 24 dB below the tone intensity. The masked threshold drops more steeply when the tone frequency is raised than when it is lowered, corresponding to the usual frequency asymmetry of auditory masking. Subject: JLH.

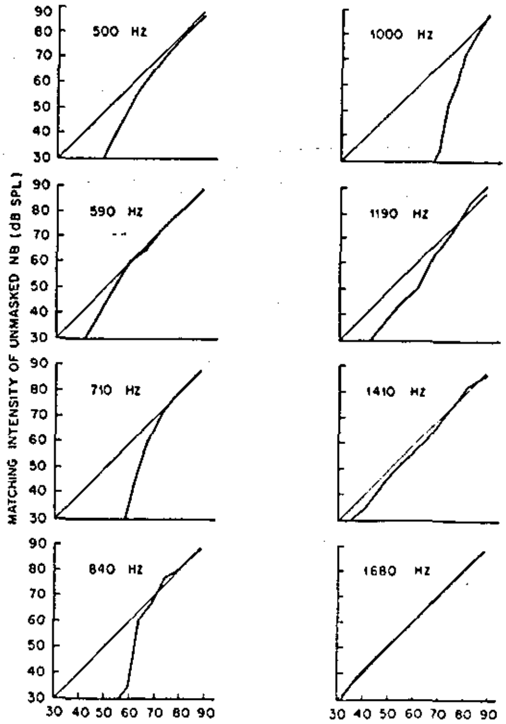
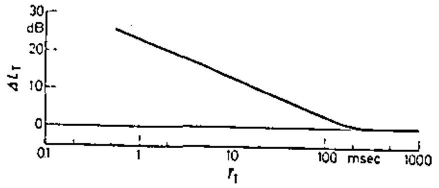


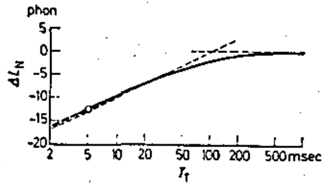
Figure 2.3-25 [19] INTENSITY OF MASKED NOISE BURST (dB SPL)

Suprathreshold loudness measurements for a critical-band noise burst centred at 1 kHz masked by a tone of intensity 80 dB SPL and frequency as indicated in each panel. Note that for a tone frequency of 1 kHz the matching intensity of the unmasked noise burst decreases with a slope of 3 dB dB when the intensity of the masked noise burst is reduced below 80 dB SPL, in agreement with Eqn (9) with $p = 2$. Subject: JLH.



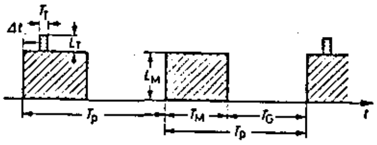
Masked threshold of single test sound impulses as a function of their duration, T_T, expressed by the level difference, $\Delta L_T = L_T(T_T) - L_T(\infty) \cdot L_T(T_T)$ is the level of the test sound impulse with the duration T_T and $L_T(\infty)$ is the level of a long sound impulse

Figure 2.3-26 [25]



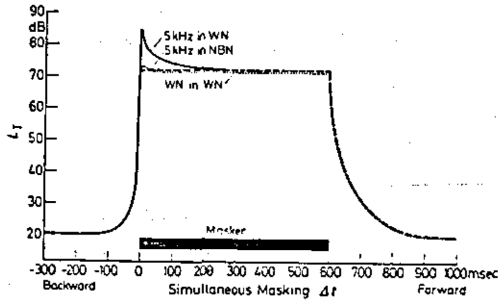
Change, $\Delta L_N = L_N(T_T) - L_N(T_T \to \infty)$, of the loudness level of a tone impulse as a function of its duration, T_T. The level from which the impulse is cut out is kept constant; the loudness level diminishes for decreasing duration. The dashed line indicates an approximation with a critical duration of 100 msec and a decrease of 10 phon when the duration is shortened by a factor of 10

Figure 2.3-27 [25]



. Sequence of test signal and masker used in measurements of temporal effects in masking

Figure 2.3-28 [25]



Backward, simultaneous and foreward masking expressed by the level, L_T, of 2 msec test signals (white noise, WN, and 5 kHz, tone-impulse) as a function of the delay time, Δt, between onset of masker and onset of signal. The level of the masker (white-noise, WN, and narrow band noise, NBN) is chosen in such a way that steady state masking is the same for all types of masker

Figure 2.3-29 [25]

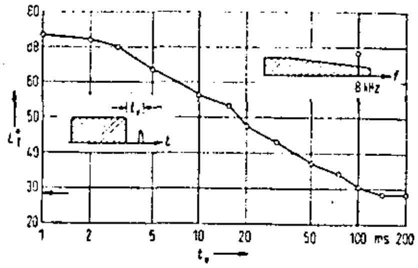
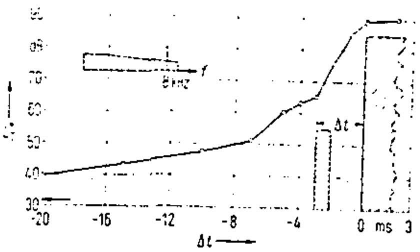


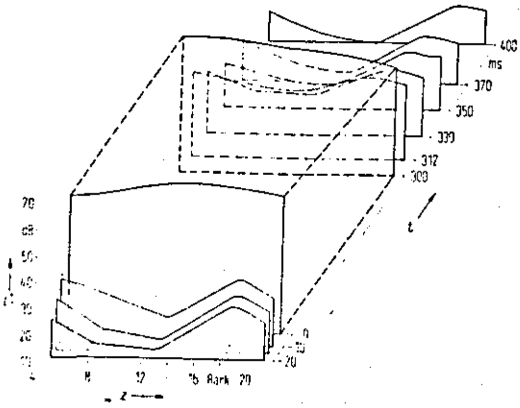
Fig. 6. Forward masking of test tone impulses by a uniform masking noise impulse as a function of delay time.
$L_M^* = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 500$ ms,
$T_T = 1$ ms, $T_{rG} = 0.5$ ms, $f_T = 8$ kHz,
arrow: threshold in quiet of test tone impulse.

Figure 2.3-30 [31]



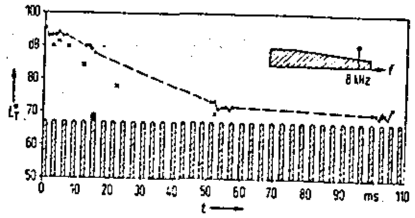
Backward masking of test tone impulses by a uniform masking noise impulse as a function of delay time.
$L^*_M = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 500$ ms,
$f_T = 8$ kHz, $T_T = 1$ ms, $T_{rG} = 0.5$ ms,
arrow: threshold in quiet of test tone impulse.

Figure 2.3–31 [31]



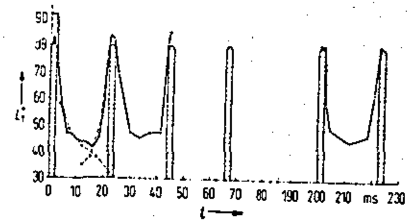
Transient masking pattern of a single uniform masking noise impulse.
$L^*_M = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 300$ ms,
$T_T = 10$ ms, $T_{rG} = 2$ ms.

Figure 2.3–32 [31]



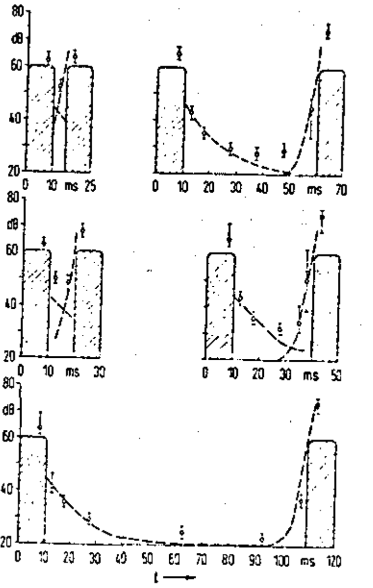
Masking of single test tone impulses by bursts of uniform masking noise.
$L^*_M = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 300$ ms,
$T_P = 500$ ms, $T_i = 1$ ms, $T_e = 2$ ms,
$f_T = 8$ kHz, $T_T = 1$ ms, $T_{rG} = 0.5$ ms,
crosses: calculated thresholds, dots: measured thresholds.

Figure 2.3–33 [31]



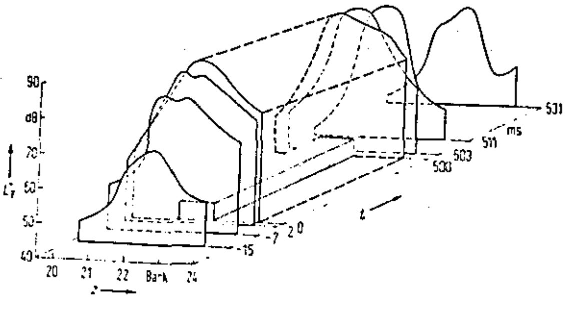
Masking of single test tone impulses by bursts of uniform masking noise.
$L^*_M = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 300$ ms, $T_P = 500$ ms,
$T_i = 2$ ms, $T_e = 20$ ms,
$f_T = 8$ kHz, $T_T = 1$ ms, $T_{rG} = 0.5$ ms,
measured thresholds: dots and solid lines, calculated thresholds: crosses and dashed lines.

Figure 2.3–34 [31]



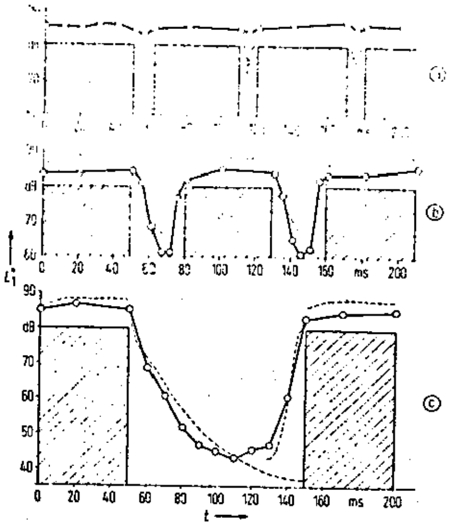
Masking of single test tone impulses by bursts of uniform masking noise.
$L^*_M = 60$ dB, $\Delta f_M = 16$ kHz, $T_M = 300$ ms, $T_P = 500$ ms,
$T_i = 10$ ms, $T_e = 5, 10, 30, 50, 100$ ms,
$f_T = 2$ kHz, $T_T = 5$ ms, $T_{rG} = 2$ ms,
circles: measured thresholds, dashed curves: estimated masking functions.

Figure 2.3–35 [31]



Transient masking pattern of a single critical band noise masker impulse.
$L^*_M = 70$ dB; $f_M = 8.5$ kHz; $\Delta f_M = 1800$ Hz;
$T_M = 500$ ms; $T_r = 1$ ms; $T_{rG} = 0.5$ ms;
hatched bar: spectral and temporal extent of masker impulse.

Figure 2.3–36 [32]



Masking of single test tone impulses by critical band noise masker bursts.
$L_M^* = 70$ dB; $f_M = 8.5$ kHz; $\Delta f_M = 1800$ Hz;
$T_M = 350$ ms; $T_P = 500$ ms; $T_i = 50$ ms;
$T_T = 1$ ms; $T_{rG} = 0.5$ ms; $f_T = 8.5$ kHz;
solid: measured masking functions;
dashed: estimated masking functions;
hatched: temporal structure of masker bursts.
(a) $T_g = 10$ ms;
(b) $T_g = 30$ ms;
(c) $T_g = 100$ ms.

Figure 2.3-37 [32]

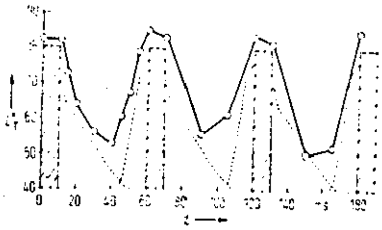
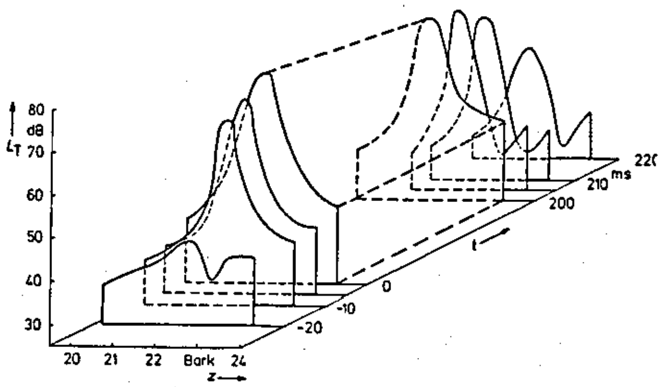


Fig. 18. Masking of single test tone impulses by a critical band noise masker burst.
$L_M^* = 70$ dB; $f_M = 8.5$ kHz; $\Delta f_M = 1800$ Hz;
$T_M = 350$ ms; $T_i = 10$ ms; $T_g = 50$ ms; $T_P = 500$ ms;
$T_T = 1$ ms; $T_{rG} = 0.5$ ms; $f_T = 8.5$ kHz;
solid: measured masking function;
dashed: estimated masking function;
batched: temporal structure of masker burst.

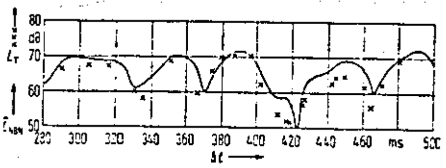
Figure 2.3-38 [32]

too.



Transient masking pattern of a sinusoidal masker impulse at 21.5 Bark
$L_M = 70$ dB; $T_M = 200$ ms; $T_T = 2$ ms; $t_a = 1$ ms

Figure 2.3-39 [30]



Threshold level L_T (×××) of 2 ms long 3-kHz-tone impulses masked by the artificial narrow band noise centered at 2 kHz with a bandwidth of 32 Hz and with the instantaneous level $\hat{L}_{NBN}$ (solid curve) as function of the delay time Δt. The "effective" level L_{NBN} of the masking noise is 70 dB.

Figure 2.3-40 [33]

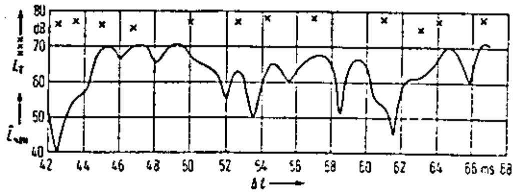
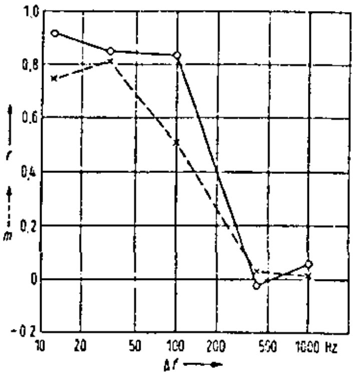


Figure 2.3-41 [33]

Threshold level L_T (×××) of 2 ms long 4-kHz-tone impulses masked by artificial narrow band noise centered at 4 kHz with a bandwidth of 1 kHz and with the instantaneous level $\hat{L}_{NBN}$ (solid curve) as function of the delay time Δt. The "effective" level of the masking noise is 70 dB.



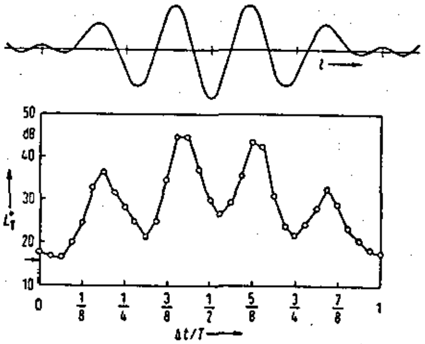
Correlation coefficient r (circles and solid line) and regression factor m (crosses and dashed line) between the instantaneous level $\hat{L}_{NBN}$ of the masking noise and the masked threshold L_T of 2-ms-tone impulses as function of the bandwidth Δf of the masking noise. Centre-frequency of the noise and tone-frequency of the impulses are each 4 kHz.

Figure 2.3-42 [33]

Average temporal distance D of envelope maxima.

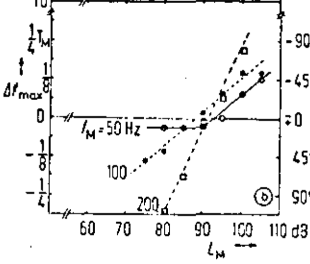
Δf Hz	$\Delta L = 0$dB D ms	$\Delta L = 3$dB D ms	$\Delta L = 6$dB D ms	$\Delta L = 10$dB D ms
10	190	252	520	3906
30	63.5	84.0	173	1302
100	19.1	25.2	52.1	390
300	6.35	8.40	17.4	130
700	2.72	3.60	7.44	55.8
1800	1.06	1.40	2.90	21.7

Table 2.3–T1 [34]



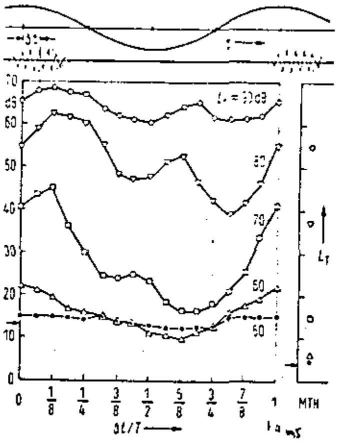
Masking period pattern of a two-tone-masker with $L_4 = 65$ dB ($f_4 = 112$ Hz) plus $L_5 = 65$ dB ($f_5 = 140$ Hz). Test signal: $f_T = 2752$ Hz; $t_i = 1.5$ ms; $t_{rG} = 0.5$ ms; octave at 2.8 kHz; $f_r = 28$ Hz. Arrow: threshold in quiet. Upper graph: time pattern of the voltage producing the masker.

Figure 2.3–44 [35]



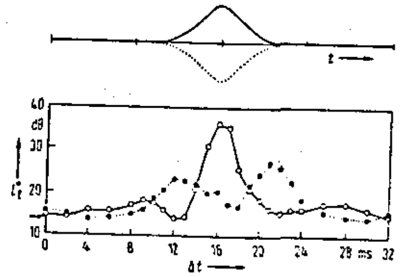
(a) Maximal level $L_{T\,max}^*$ within the masking-period patterns and (b) temporal occurrence Δt_{max} of this value within the period as function of the masker level L_M. Parameter is the frequency f_M of the masker. For details, see text.

Figure 2.3–46 [36]



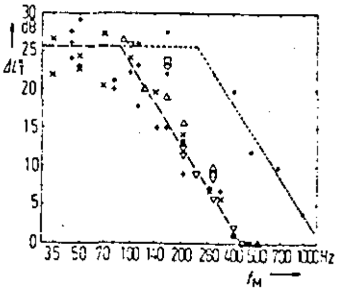
Masking period pattern of a single 70 Hz-masker, the level L_1 of which is parameter. The test signal level L_T^* of the 2 kHz-pulse is plotted as function of the time difference Δt as indicated in the two top drawings. Test signal: t_i 2.5 ms; t_{rG} 1 ms; octave band filter at 2 kHz, repetition rate f_r = 70 Hz. Right ordinate: masked threshold (L_T) produced with continuous test tones.

Figure 2.3–43 [35]



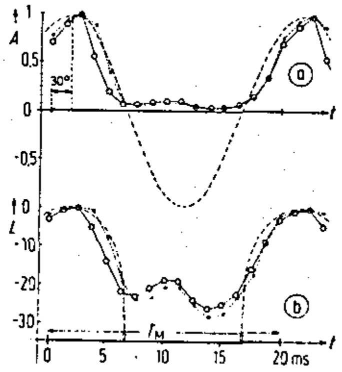
Masking period pattern of a Gaussian DC-pulse-masker ($t_i = t_{rG} = 3$ ms) with a peak level $\hat{L} = 80$ dB in normal (solid) and reversed (dotted) condition. Test signal: $f_T = 2840$ Hz; $t_i = 1.5$ ms; $t_{rG} = 0.5$ ms; octave at 2.8 kHz; $f_r = 31.2$ Hz. Upper graph: time patterns of the voltages producing the maskers.

Figure 2.3–45 [35]



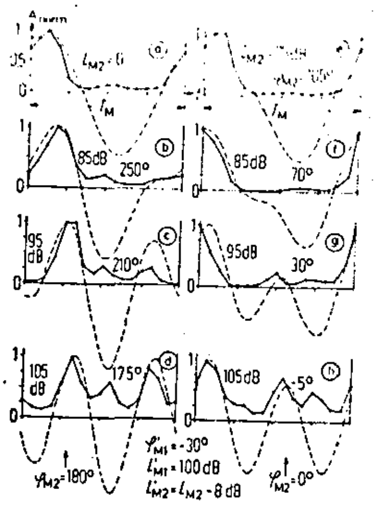
Difference ΔL_T^* between maximum and minimum of masking-period patterns elicited by pure tone maskers of different frequencies f_M under optimal conditions (cf. paragraph 3). Dashed lines: approximation of data produced by five subjects. Dotted lines: approximation of data of one subject with different behavior.

Figure 2.3–47 [36]



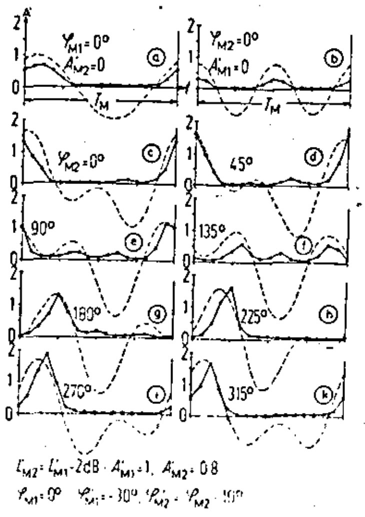
Comparison between masking period pattern elicited by a 50-Hz, 100-dB masker (see Figs. 1 and 5) and time function of the 50-Hz tone with corresponding peak value, shifted by $\varphi'_{M1} = 30°$. (a) linear ordinate scale as used in neurophytologically measured period histograms. (b) logarithmic ordinate scale as used in psychoacoustically measured masking period patterns. Circles: averaged data from 11 measurements on the same subject obtained throughout a period of one year. Dots: data from Fig. 5.

Figure 2.3-48 [37]



Masking period patterns (dots) shown in Fig. 1, replotted on linear ordinate scale. Parameter is the level L_{M2} of the second harmonic (100 Hz) as indicated. For comparison, calculated time functions based on the level relation of the masker components and on the phase shifts φ'_M (as indicated) of the patterns elicited by single masker components are drawn as dashed lines. The patterns as well as the time functions are normalized for each diagram separately.

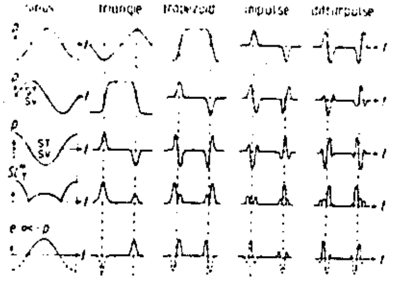
Figure 2.3-49 a [37]



$L_{M2} = L_{M1} - 2 dB \cdot A'_{M1} = 1, A'_{M2} = 0.8$
$\varphi_{M1} = 0° \quad \varphi'_{M1} = -30°, \varphi'_{M2} = \varphi_{M2} - 90°$

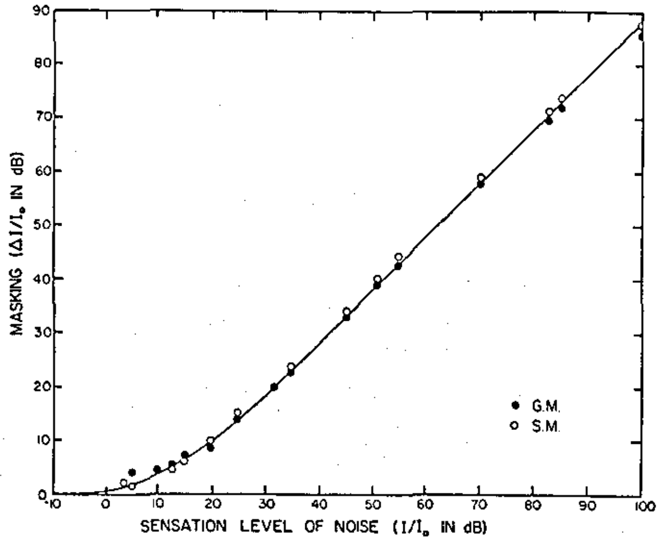
Same as Fig. 15, but from data shown in Fig. 5. The amplitudes are normalized only once for all patterns and calculated time functions (based on the data given at the bottom). Parameter is the phase angle φ_{M2} of the second harmonic of the masker, as indicated.

Figure 2.3-49 b [37]



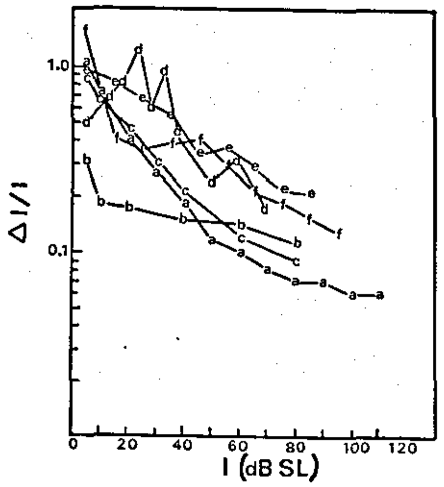
Summarized presentation of idealized time functions of the sound pressure p, of its first ($\dot{p}$) and second ($\ddot{p}$) derivative, of the MPP produced by $p(t)$ (represented by the sensation level $SL_{\ddagger}$ of the test tone burst) and of the time variant psychoacoustical excitation e which, according to the model, is proportional to $-\ddot{p}$ at frequencies below 40 Hz. St and SV indicate the displacement of the basilar membrane towards scala tympani and scala vestibuli, respectively.

Figure 2.3-50 [38]



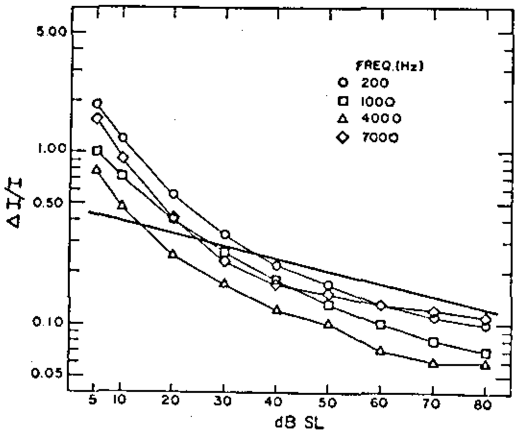
The just noticeable increment in the intensity of a noise, ΔI, as a function of the intensity of the noise, I. The solid straight line represents the relation $\Delta I/I = k$, or Weber's law.

Figure 2.3-51 [39]



The Weber fraction as a function of SL (for 1000 Hz) from several studies: (a) Riesz , (b) and (c) Harris : (d) McGill and Goldberg ; (e) Campbell and Lasky : (f) Luce and Green .

Figure 2.3-52 [41]



The Weber fraction ΔI/I as a function of SL. Riesz's data are shown by the curves with symbols; the straight line shows the results of Jesteadt et al.

Figure 2.3-53 [42]

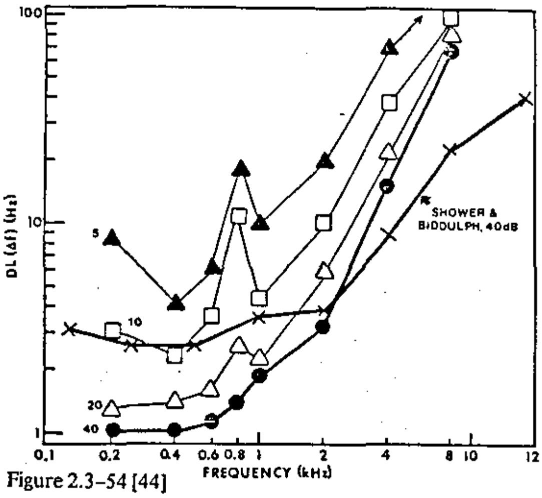
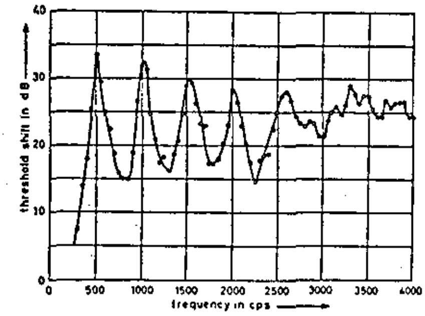


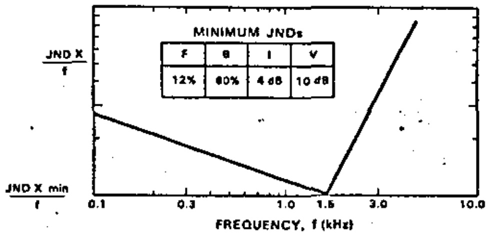
Figure 2.3-54 [44]

Frequency DL as a function of frequency at SLs of 5 dB (filled triangles), 10 dB (open squares), 20 dB (open triangles), and 40 dB (filled circles): based on data of Wier et al. . Shower and Biddulph's FM data at 40 dB SL (X's) are shown for comparison.



Masking of a tone pulse of 20 msec presented immediately after a 200-msec pulse of a complex tone consisted of the first 12 harmonics of 500 cps. The points indicate the average values of 4 observers.

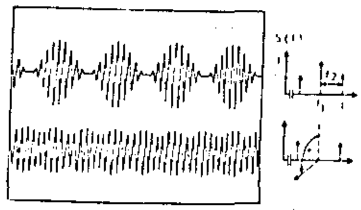
Figure 2.3-56 [45]



MINIMUM JNDs			
F	B	I	V
12%	60%	4 dB	10 dB

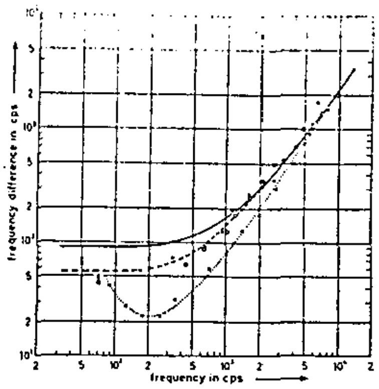
Estimated values of the formant parameter JND's as a function of frequency. *X* represents the formant parameters: center frequency *F*, bandwidth *B*, peak power *I*, and valley power *V*.

Figure 2.3-58 [49]



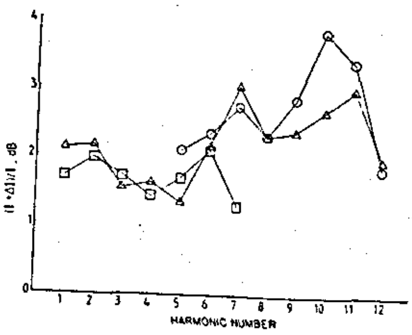
AM signal (*top*) is transformed to a QFM signal (*bottom*) by a phase change. The AM signal contains only amplitude modulation, whereas the QFM is both amplitude- and frequency-modulated. For the signals shown, $f_1/f_2 = 10$ and the AM is modulated 100%.

Figure 2.3-60 [52]



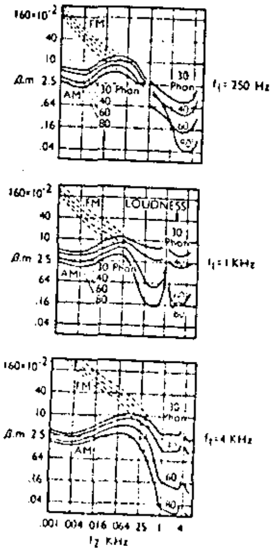
Frequency difference between the partials of complex sounds required to hear them separately. The open points represent the values for complex tones (Fig. 6), the solid points for inharmonic complexes of tones (Fig. 8), and the crosses for two-tone signals (Fig. 10). The solid curve represents the critical bandwidth as a function of frequency, after Zwicker, Flottorp, and Stevens. The dashed curve and the dotted curve fit the data points concerning multitone and two-tone signals, respectively.

Figure 2.3-55 [45]



Results from experiment III, averaged across subjects. Intensity DLs for individual harmonics are plotted as a function of harmonic number for complex tones with harmonics as follows: 1–7 (squares); 1–12 (triangles); 5–12 (circles).

Figure 2.3-57 [46]



Modulation thresholds for sinusoidally modulated AM and FM for various carrier frequencies. β: FM modulation index m: AM modulation depth.

Figure 2.3-59 [52]

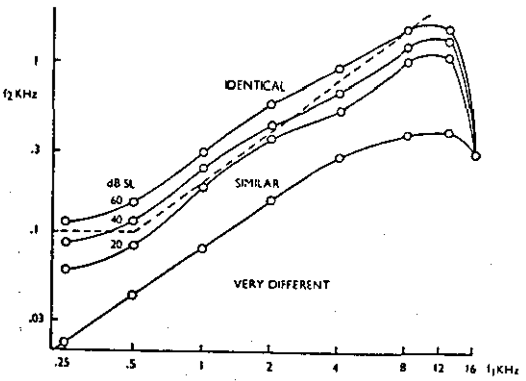
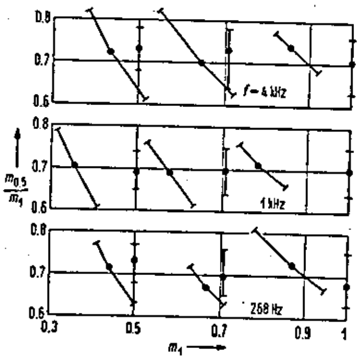


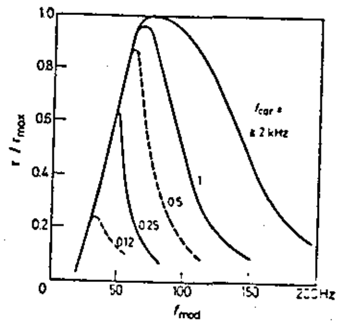
Figure 2.3-61 [52]

Comparisons of AM and QFM (100% modulation) as a function of carrier frequency f_1, modulation frequency f_2, and carrier level. The results of the four subjects were logarithmically averaged for each of the carrier levels 20, 40, and 60 dB SL.. (At the carrier 250 Hz, the carrier levels were 15, 31, and 47 dB SL, corresponding approximately to 20, 40, and 60 phon.) The curves suggest some relation to the critical-band scale (---), but there is a clear dependence upon carrier level.



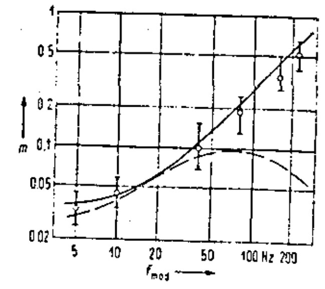
The ratio $m_{0.5}/m_1$ which corresponds to the roughness ratio 0.5: 1, as a function of m_1.
$m_{0.5}$, m_1: degrees of modulation of sinusoidally amplitude modulated tones (AM tones); f: carrier frequency.
Resulting points with vertical bars: criterion "half as rough", remaining points: criterion "twice as rough".
The bars represent the interquartiles (25% and 75% of the answers, respectively).

Figure 2.3-62 [54]



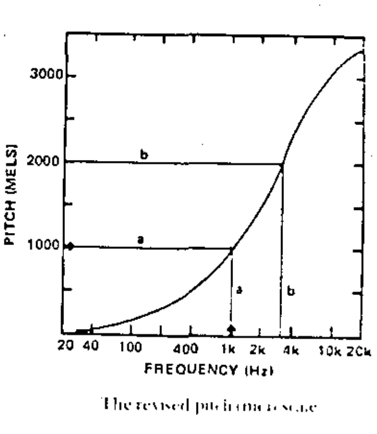
Relative roughness, $\frac{r}{r_{max}}$, as a function of the modulation frequency, f_{mod}, for amplitude modulated pure tones with different carrier frequencies, f_{car}

Figure 2.3-63 [25]



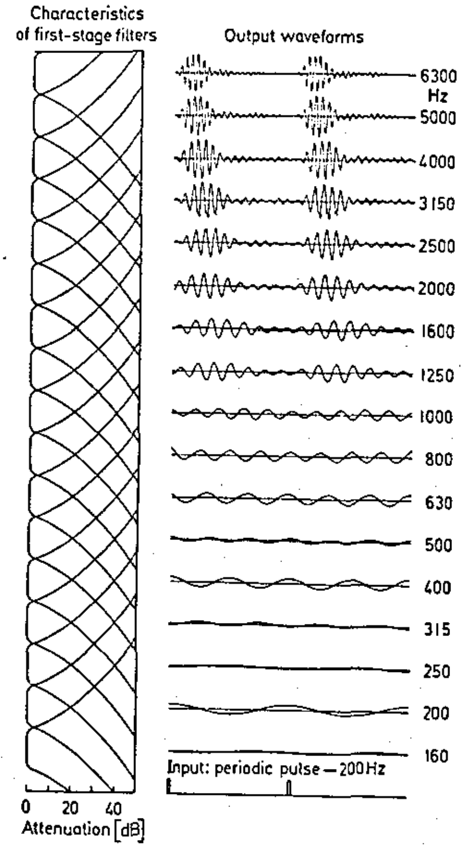
Open circles: Degree of modulation m of an AM tone corresponding to just noticeable fluctuations ($f_{mod} < 20$ Hz) or roughness ($f_{mod} > 20$ Hz), as a function of f_{mod}.
f = 1 kHz; SPL 60 dB.
Dashed line: AM-threshold after ZWICKER
Solid line: Approximation of the results by a low-pass RC-network with a time constant of 13 ms.

Figure 2.3-64 [54]



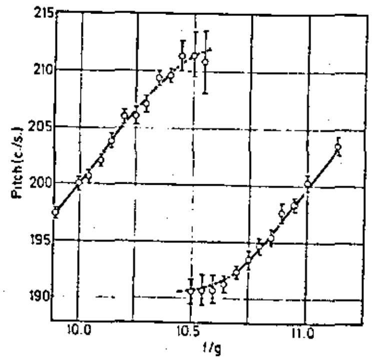
The revised pitch (mel) scale

Figure 2.3-65 [56]



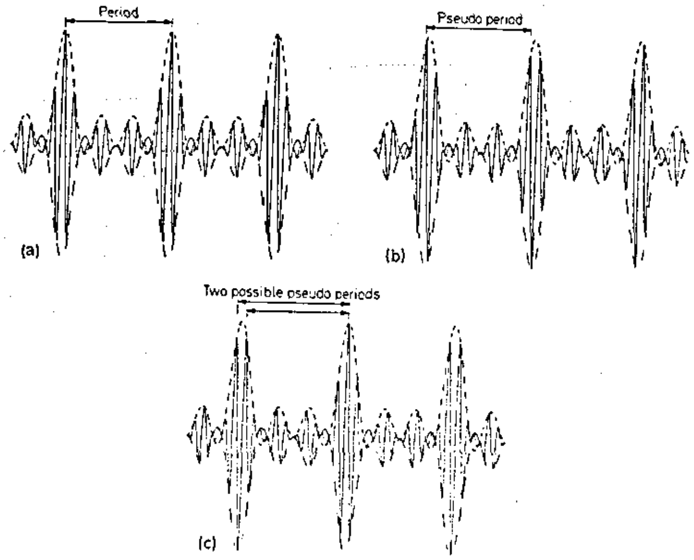
Periodicity of signals and aural resolution of components: a more modern representation.

Figure 2.3-66 [58]



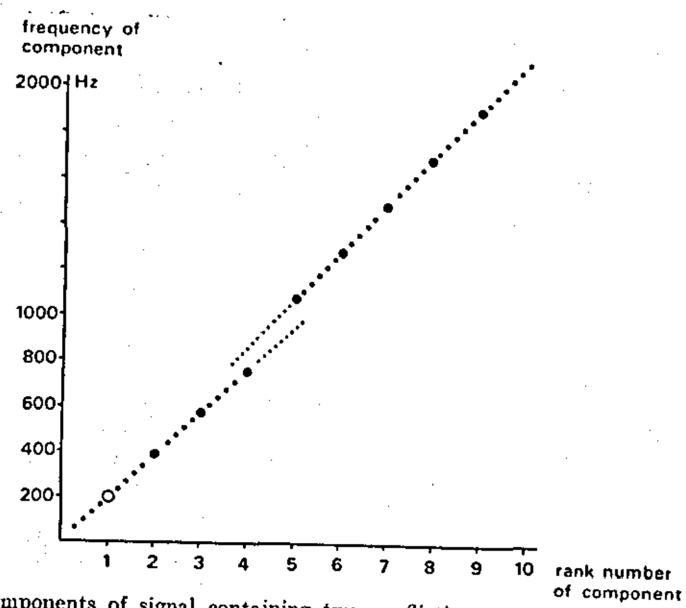
Pitch of inharmonic signals as measured by DE BOER

Figure 2.3-67 [58]



Generalization of the concept of "period": (a) Purely periodic signal (waveform). (b) Inharmonic signal: the envelopes (shown by dashed and dotted lines) are periodic but the signal is not periodic. The "pseudo-period" denotes the time distance over which the signal repeats itself approximately. (c) Signal containing only odd-numbered harmonics. Two pseudo-periods are possible

Figure 2.3-68 [58]



Components of signal containing two conflicting types of information for pitch. The slopes of the dotted lines indicate the frequencies f_0 and $f_0 + \Delta f$. Only the A-signal is shown. The fundamental of the lower band is indicated by an open circle. Note that both dotted lines meet at the origin

Figure 2.3-69 [58]

1. J.C. Stevens and Miguelina Guirao "Individual Loudness functions" J. Acoust. Soc. Am. Vol 36, No 11, Nov. 1964 pp2210-2213

2. R.P. Hellman and J. Zwislocki "Some factors affecting the estimation of Loudness" J. Acoust. Soc. Am. Vol. 33, No 5, May 1961, pp687-694.

3. J.P.A. Lochner and J.F. Burger "Form of the Loudness Function in the Presence of Masking Noise" J. Acoust. Soc. Am., Vol. 33, No. 12, Dec. 1961, pp1705-1707.

4. R.P. Hellman and J. Zwislocki "Loudness function of a 1000-cps tone in the presence of a masking noise" J. Acoust. Soc. Am., Vol. 36, No. 9, Sept. 1964 pp1618-1627.

5. Fletcher, H. "Speech and hearing in Communication" D. Van Nostrand Company; Princeton New Jersey, (1958).

6. R.H. Ehmer, "Masking Patterns of Tones", J. Acoust. Soc. Am. Vol. 31, No 8, August, 1959, pp1115-1120.

7. E. Zwicker "On a psychoacoustical equivalent of tuning curves" In: facts and Models in Hearing, E. Zwicker and E. Terhardt (eds.) Springer-Verlag, Berlin Heidelberg New York 1974 pp132-141.

8. L.L.M. Vogten "Pure-tone Masking; A New result from a new method" in: Facts and Models in Hearing, E. Zwicker and E. Terhardt (eds.) Springer-Verlag, Berlin Heidelberg New York 1974 pp142-155.

9. R.H. Ehmer "Masking by Tones vs Noise bands", J. Acoust. Soc. Am. Vol. 31, No. 9, 1959 pp1253-1256.

10. E. Zwicker, G. Flottorp, S.S.Stevens "Critical Bandwidth in Loudness Summation", J. Acoust. Soc. Am. Vol 29, No 5, May 1957, pp548-557.

11. D.D. Greenwood "Auditory Masking and the Critical Band" J. Acoust Soc. Am. Vol 33, No. 4, April, 1961.

12. E. Zwicker "Subdivision of the Audible Frequency Range into Critical Bands" J. Acoust. Soc. Am. Vol. 33, No. 2, Feb. 1961 (p248).

13. D.D. Greenwood "Critical bandwidth and the frequency coordinates of the Basilar Membrane" J. Acoust. Soc. Am. Vol. 33, No. 10, 1961, pp1344-1356.

14. B. Scharf "Loudness Summation under Masking" J. Acoust. Soc. Am. Vol. 33, No. 4, April 1961, pp503-511.

15. B. Scharf "Partial Masking" Acoustica Vol. 14, 1964 pp16-23.

16. B. Scharf "Patterns of Partial Masking" In: Seventh International Congress on Acoustics, Budapest 1971, 23H6 pp461-464.

17. R.P. Hellman "Asymmetry of masking between noise and tone" Perception and psychophysics 1972 Vol. 11(3) pp241-246.

18. M.R. Schroeder, B.S. Atal, J.L. Hall, "Optimizing digital Speech coders by exploiding masking properties of the human ear" J. Acoust. Soc. Am. 66(6), Dec. 1979 pp1647-1652.

19. M.R. Schroeder, B.S. Atal and J.L. Hall "Objective Measure of Certain Speech Signal Degradations Based on Masking Properties of Human Auditory Perception" In: Frontiers of Speech Communication Research, B. Lindblom and S. öhman (eds.) Academic Press, 1979 pp217-229.

20. Young, I.M. & Wenner, C.H. Masking of white noise by pure tone, frequency-modulated tone, and narrow band noise. J. Acoust. Soc. Am. 1967, 41, 700-706.

21. E. Zwicker "Temporal Effects in Simultaneous Masking and Loudness" J. Acoust. Soc. Am. 38 pp132-141 1965.

22. J. Zwislocki "Theory of Temporal Auditory Summation" J. Acoust. Soc. Am. Vol. 32, No. 8, 1960 pp1046-1060.

23. E.C. Sheeley and R.C. Bilger "Temporal Integration as a function of Frequency" J. Acoust. Soc. Am. Vol. 36, No. 10, pp1850-1857 1964.

24. E. Zwicker "Temporal effects in Psychoacoustical Excitation" in: Basic Mechanisms in Hearing" A.R. Moller (Ed.) Academic Press 1973 pp809-827.

25. E. Zwicker "Scaling" In: Handbook of Sensory Physiology V/2, Auditory System, W.D. Keidel and W.D. Neff eds. Springer-Verlag Berlin Heidelberg New York 1975, p401-448.

26. H. Fastl "Masking Patterns of Subcritical versus Critical Band-Maskers at 8.5 KHz" Acustica Vol. 34 (1976), p167-171,

27. E. Zwicker "Temporal effects in simultaneous Masking by White Noise Bursts" J. Acoust. Soc. Am. Vol. 37, No. 4, 1965, pp653-663.

28. E. Zwicker and H. Fastl "On the development of the citical band" J. Acoust. Soc. Am. Vol. 52, No.(2) (part 2) 1972 pp699-702.

29. H.N. Wright "Backward masking for tones in narrow-band noise" J. Acoust. Soc. Am. Vol. 36, No. 11, pp2217-2221.

30. Fastl, H. "Transient Masking Pattern of Narrow Band Maskers" in: Facts and Models in Hearing" E. Zwicker and E. Terherdt eds. Springer-Verlag 1974 pp251-257.

31. Fastl, H. "Temporal Masking Effects: I. Broad Band Noise Masker" Acustica Vol. 35, No. 5, 1976 pp287-302.

32. Fastl, H. "Temporal Masking Effects: II. Crilcal Band Noise Masker" In Acustica Vol. 36, No. 5, 1976/77 pp317-331.

33. E. Zwicker and H. Schütte "On the time pattern of the threshold of Tone impulses masked by Narrow Band Noise" Acustica Vol. 29, 1973 pp343-347.

34. H. Fastl "Loudness and Masking Patterns of Narrow Noise Bands" Acustica, Vol. 33, 1975 pp267-271.

35. E. Zwicker "Influence of a Complex Masker's Time Structure on Masking" Acustica Vol. 34 (1976) pp138-146.

36. E. Zwicker "Psychoacoustic equivlent of period histograms" J. Acoust. Soc. Am. Vol. 59, No. 1, Jan. 1976. pp166-175.

37. E. Zwicker "Masking Period Patterns of harmonic complex tones" J. Acoust. Soc. Am. Vol. 60 (No. 2) Aug. 1976, pp429-439.

38. E.Zwicker "Masking-period patterns produced by very low frequecy maskers and their possible relation to basilar membrane displacement" J. Acoust. Soc. Am. Vol.(61) No. 4, April 1977. pp1031-1040.

39. G.A. Miller "Sensitivity to changes in intensity of white noise and its relation to masking and loudness" J. Acoust. Soc. Am. 1947, 14, pp609-619.

40. R.R. Reisz "Differential intensity sensitivty of the ear for pure tones". Physio. Rev. 31, pp867-875 1928.

41. R.D. Luce and D.M. Green "Neural coding and Physiological discrimination data"J. Acoust. Soc. Am. 56, 1554-1564, 1974.

42. W. Jesteadt, C.C. Wier and D.M. Green "Intensity discrimination as a function of frequency and sensation level" J. Acoust. Soc. Am. 61, 169-177, 1977.

43. E.G. Shower and R. Biddulph "Differential pitch sensitivity of the ear" J. Acoust. Soc. Am. 3, 275-287, 1931.

44. C.C. Wier, W. Jesteadt and D.M. Green "Frequency discrimination as a function of frequency and sensation level" J. Acoust. Soc. Am. 61, 178-184, 1977.

45. R. Plomp "The ear as a frequency analyser" J. Acoust. Soc. Am. Vol. 36, 9. Sept. 1964 pp1628-1636.

46. B.C. J. Moore, B.R. Glasberg, M.J. Shailer "Frequency and intensity difference linens for harmonics within complex tones" J. Acoust. Soc. Am. 75(2), Feb. 1984 pp550-561.

47. Flanagan J.L. "Speech analysis sythesis and perception" springer-Verlag, Berlin Heidelberg - New York 1972.

48. D.H. Klatt "Discrimination of fundamental frequency contours in synthetic speech: Implications for models of pitch perception" J. Acoust. Soc. Amer. Vol. 53, pp8-16, 1973.

49. O. Ghitza and J.L. Goldstein "Scalar LPC Quantization based on Formant JNDs" IEEE trans. on Acoust. Speech Signal Proc. Vol. ASSP-34, No. 3, 1986 pp697-708.

50. R. Viswanathan, J. Makhoul, "Quantization properties of transmission parameters in linear predictive systems" IEEE Trans. Acoust. Speech, Signal, Processing, Vol. ASSP-23, p309-321, June, 1975.

51. R.C. Mathes and R.L. Miller "Phase effects in Monaural Perception" J. Acoust. Soc. Am. Vol. 19, No. 5, Sept. 1947, pp780-797.

52. J.L. Goldstein "Auditory Spectral filtering nd Monaural Phase Perception" J. Acoust. Soc. Am. Vol. 41 (No. 2) 1967 pp458-479.

53. M.R. Schroeder "Models of Hearing" IEEE proc. Vol. 63, No. 9, Sept. 1975, pp1332-1350.

54. E. Terhardt "On the perception of Periodic Sound Fluctuations (Roughness)" Acustica Vol. 30, 1974 pp201-213).

55. A Vogel "Roughness and its relation to the time-pattern of psychoacoustical excitation" in: "Facts and Models of hearing" E. Zwicker and E. Terhardt (eds.) Springer-Verlag, 1974 pp241-250.

56. S.S. Stevens and J. Volkmann, The relation of pitch to frequency: A revised scale, Amer. J. Psych. 53 pp329-353 (1940).

57. B. Scharf "Critical Bands" In: Foundations of Modern Auditory theory (J.V. Tobias Ed.) Vol. I, Academic Press, 1970 pp157-202.

58. E. de Boer "On the "residue" and Auditory Pitch Perception" In: Handbook of sensory Physiology Vol. V/3 Springer-Verlag, 1976 pp479-583.

59. J. Makhoul and M. Berouti "Predictive and residual encoding of Speech" J. Acoust. Soc. Am. 66(6) Dec. 1979 pp1633-1641.

60. E. de Boer "Pitch of inharmonic signals" Nature (Lond.) 178, 535-536, 1956.

61. R.J. Rotsma "Frequencies dominant in the perception of the pitch of complex sounds" J. Acoust. Soc. Am. 1967 pp191-198.

62. R. plomp and H.J.M. Steeneken "Effect of phase on the Timbre of Complex Tones. J. Acoust. Soc. Am. Vol. 46, No. 2 (part 2) 1969.

63. G. von Bismarck "Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes" Acustica Vol. 30, 1974 pp146-158.

64. G. Von Bismarck "Sharpness as an Attribute of the Timbre of Steady Sounds" Acustica Vol. 30, 1974 pp159-172.

65. L.C.W. Pols, L.J.T. van der Kamp, R. Plomp "Perceptual and Physical Space of Vowel Sounds" J.

Acoust. Soc. Am. Vol. 46, No. 2 (part 2) 1969 pp458-467.

66. W. Klein, R. Plomp, L.C.W. Pols "Vowel Spectra, Vowel Spaces, and Vowel identification" J. Acoust. Soc. m. Vol. 48, No. 4, pp999-1009, 1970.

67. L.C.W. Pols, H.R.C. Tromp, R. Plomp "Frequency analysis of Dutch vowels from 50 male speakers" J. Acoust. Soc. Am. Vol. 53, No. 4, 1973 pp1093-1101.

a) D.M. Green, "An Introduction to Hearing", Lawrence Erlbaum Associates; Hillsdale, New Jersey (1976).

b) H. Fletcher, "Speech and Hearing in Communication", D. Van Nostrand Company'; Princeton New Jersey (1958).

c) M.R. Schroeder, "In Recognition of Complex Acoustic Signals", Life Sciences Research, Report 5, edited by T.H. Bullock (Dahlem Konferenten) (Abakon Verlag, Berlin) pp323-328 (1977).

d) B.C.J. Moore "An Introduction to the Psychology of Hearing", Academic Press, London (1982).

e) I.C. Whitfield, "The Auditory Pathway", Edward Arnold, London (1967).

f) J.L. Hall, "In Hearing Research and Theory", Volume 1, edited by J.V. Tobias and Earl D. Schubert, Academic Press, London (1981).

2.4 Models of psychophysical phenomena

Many models have been devised throughout the years to explain and predict psychophysical responses. Early models were concerned with the intelligibility of speech signals and the closely related Articultion Index [6, 3, 4, 5, 6]. The degredations studied were usually due to filtering or noise masking. These first experiments helped to establish acceptable noise conditions for telephone circuits as well as the bandwidth of the telephone band, 300 Hz - 3.4 KHz.

Other models attempted to predict specific phenomena such as masking [b] temporal summation (integration) [1,2], loudness calculation [7, 8, 9], Monaural phase perception [10], backward masking [11, 12] or pitch [13].

Some of them made use of physiological phenomena [1, 2, 10, 11, 12] such as the probabilistic nature of the neural transducer whilst others builded upon strictly psychophysical phenomena. Nearly all of them made use of some mechanisms of auditory selectivity. In the remaining parts of this section we will concentrate on a model which aims to encompass all the known psychophysical phenomena and explain them through the use of only a limited number of concept elements. These concepts are the critical band, the excitation pattern and some recent time related elements which form the excitation critical band rate-time pattern. This model can be used to explain and predict masking jnds loudness, pitch, timbre, periodicity pitch, roughness and other phenomena. It is mainly due to Eberhard Zwicker and was originally proposed to predict loudness summation and masking but developed in the last 20 years into a model able to describe the multitude of phenomena mentioned above without losing its conceptual simplicity. A review on the critical band can be found in [14].

2.4.1 The Critical band scale

Analytical expressions have been given to transform from a frequency scale to a critical band scale. The equivalent unit of measurement on a critical band scale is the Bark. Some values are given in table (2.4-T1). For frequency f in KHz, critical-band rate Z_c in Bark and arctan() in radians the following expression has been proposed [15].

$$Z_c = 13.0 \arctan (0.76f) + 3.5 \arctan \left(\frac{f}{7.5}\right)^2 \qquad 2.4\text{-}1$$

which transforms the linear frequency scale into the Bark scale.

The critical bandwidth CB_c is given by [15]

$$CB_c = 25.0 + 75.0 [1 + 1.4(f)^2]^{0.69} \qquad 2.4\text{-}2$$

where f is in KHz. The above are applicable over the whole audible frequency range.

Another set of expressions can be found in [16] which provides good fit for frequencies less than 5 KHz.

$$f = 650 \sinh (Z_c/7) \qquad 2.4\text{-}3$$

$$CB_c = (650/7) \cosh (Z_c/7) \qquad 2.4\text{-}4$$

where f is the frequency in Hz and Z_c the critical band number (or rate) in Bark.

A rule of thumb is that the critical band is of width of 100 Hz below 500 Hz and one-sixth of the center frequency above that. A plot of critical band rate as a function of frequency is shown in figure [2.4-1]. It is generally accepted that the critical bandwidth is related to a filter with finite slopes. The filter shape was "predicted" from masking patterns (audiograms) and is shown schematized in figure [2.4-2].

These patterns were derived by Zwicker through masking experiments. The frequency difference between the two 3-dB points corresponds to the critical bandwidth [17]. The slopes of the patterns towards lower frequencies are independent of center frequency and level and have a steepness of about 27 dB/Bark (27 dB/critical band). The slopes of the patterns towards higher frequencies depend clearly on level [17]. It is interesting to note that first, a critical bandwidth was calculated from psychophysical measurements and then, this bandwidth assigned to this filter whose shape was determined form masking audiograms. Some authors argued that the shape of the auditory filter would affect the value of the critical bandwidth in the first place: For the same psychophysical measurements, different sets of bandwidths and shapes could be used to account for the data [18]. Some of the experiments disputing the shapes of the filters and the size of the critical bands are very recent [19, 20, 21, 22, 23, 24, 25] which goes to show how little is known about the psychophysics of auditory system, indeed about its most basic characteristic, that of the critical band, and therefore the difficulty of constructing models with predictive powers with any degree of certainty.

2.4.2 A model of loudness summation

The model of loudness summation by Zwicker and Scharf [26] will now be described. Its aim is to permit the calculation of loudness from the physical spectrum in the presence of masking noise. Figure 2.4-3a,b traces the transformations that are assumed to occur when the ear is stimulated by pure tone [2.4-3a] and by white noise [2.4-3b]. The physical spectra are shown on the top. The first stage is the envelope of stimulation of the BM (second from the top). It can be seen that even the pure tone with a line spectrum produces a displacement over a wide area of the BM. The next step is the excitation pattern and represents the level of neural activity due to the displacement of the BM. It can be seen that this is much narrower than the BM displacement curve. (The excitation pattern is derived from masking data). Note that from here on values along the horizontal axis are plotted against the critical band rate or tonalness (Z). The excitation pattern is then transformed into a loudness pattern through an equation relating Specific Loudness N' to Excitation E. Finally, the total loudness produced by the original stimuli is the integral of the area under the specific loudness pattern. Note that, in the case of white noise, the part of the noise spectrum in the upper frequencies contributes more to loudness than the part in the lower frequencies.

The loudness of the noise can be reduced if it is suitably shaped so that it has a lowpass spectrum. We will now look at the transformations taking place in more detail. The excitation patterns (and loudness patterns) which are assumed to correspond to activity in the nervous system are derived from masking patterns. This idea is not new (b). The masking pattern for a given sound is the plot of the masked threshold of every narrow band stimulus as a function of its frequency. It is not necessary to measure masked thresholds of every sound. A set of standard masking patterns and their derived excitation patterns can be used which represent the masking patterns of any sound narrower than or equal to a critical band (but remember the effects of envelope fluctuations for very narrow bands). Combinations of patterns centered at different frequencies can represent the masking patterns of sounds wider than a critical band

[27]. A set of such masking patterns are shown in figure 2.4-4. It is then necessary to determine from the masked threshold the excitation level within the nervous system produced by the masking sound. The two can be related through self masking or the intensity jnd. Since in such measures the excitation patterns of the masked noise ΔI and the masking noise I are almost identical, their ratio $\Delta I/I$ may be taken to express the ratio between the excitation of the just masked stimulus and that of the masking stimulus. The range of the intensity jnd is small and can be taken to be between 1/2 (3 dB) and 1/4 (6 dB). Therefore the minimum excitation required to mask completely is twice to four times the intensity at the masked threshold. Hence the addition of 3 - 6 dB to the masked threshold gives the value for the excitation level. The excitation level can be expressed in dB as $L_E = 10\log_{10} (E/E_o)$ where E_o is a reference value corresponding to $I_o = 10^{-16}$ watt/cm^2. The masking patterns of (fig. 2.4-4) can now be replotted as excitation patterns (figure 2.4-5). Note the change from frequency to critical band rate on the abscissa.

Although the excitation patterns shown in figure (2.4-5) are for a critical band of noise centered at 1200 Hz and having the SPL shown as the parameter on the curve, excitation patterns for subcritical bands of noise (or tones) at higher or lower frequencies are similar (by virtue of the similarity of their masking patterns). The next step is to convert from excitation E to specific loudness N'. The psychophysical equation expressing this relation is based upon Stevens's Power law [28], which in one form, says that equal intensity ratios yield equal loudness ratios. It is also based on the assumption that the loudness of any sound is the integral of the specific loudness over the Z (bark) scale. In this sense every sound involves the summation of loudness, since as seen in figure [2.4-3a] even a pure tone produces an excitation pattern which spreads over a considerable portion of the Z scale.

Stevens law may be expressed as:

$$\frac{\Delta N}{N} = K' \frac{\Delta I}{I} \qquad 2.4\text{-}5$$

where I is the intensity of a tone and N is the loudness. In terms of excitation we can assume that the equation applies over a small region on the Z scale where the excitation can be assumed to be constant. For this small region

$$\frac{\Delta N'}{N'} = k \frac{\Delta E}{E} \qquad 2.4\text{-}6$$

where N' is now the "specific" (or incremental) loudness and E the excitation of a particular location on the Z scale.

Near threshold where intensity discrimination is poor an adjustment term must be added to the denominators of the above equation. This constant Egr can be thought to represent the excitation produced by the ear by an inaudible physiological background noise. This excitation can suppress a weak excitation produced by an external stimulus thereby setting a lower limit, the absolute threshold, for the ear's sensitivity. The corresponding inaudible specific "loudness" is N'gr

$$\frac{\Delta N'}{N' + N'gr} = K \frac{\Delta E}{E + Egr} \qquad 2.4\text{-}7$$

Treating the above equation as a differential equation and integrating, we have

$$\log (N'gr + N') = k\log (Egr + E) + \log c \qquad 2.4\text{-}8$$

or

$$N'gr + N' = C\ (Egr + E)^{k} \qquad 2.4\text{-}8a$$

C can be calculated from N' = 0 for E=0 giving

$$C = \frac{N'gr}{Egr^{k}}\ , \qquad 2.4\text{-}9a$$

$$N'gr = C\ (Egr)^{k} \qquad 2.4\text{-}9b$$

The evaluation of the constant Egr depends upon the same assumption used to convert from masking to excitation patterns, namely, that the masking excitation must be twice to four times the excitation produced by the just masked tone. It is assumed that the internal background excitation is twice (to four times) the excitation E_t produced by an external tone at the absolute threshold

$$Egr = 2\ E_{t} \qquad 2.4\text{-}10$$

Using equations 2.4-9 and 2.4-10 to substitute in 2.4-8:

$$N' = N'gr\ [\ (\frac{E}{2E_{t}} + 1)^{k} - 1] \qquad 2.4\text{-}11$$

in order to express the value N'gr in relative values of excitation E_t/E_o the reference value $N'gr_o$ is introduced. With $2E_t$ replacing Egr in equation 2.4-9a we obtain

$$\frac{N'gr}{N'gr_{o}} = \left(\frac{2E_{t}}{E_{o}}\right)^{k} \quad \text{or} \qquad 2.4\text{-}12a$$

$$N'gr = N'gr_{o}(\ \frac{2E_{t}}{E_{o}}\ )^{k} \qquad 2.4\text{-}12b$$

Equation 2.4-11 may now be written

$$N' = N'gr_o \left(\frac{2E_t}{E_o} \right)^k x \left[\left(\frac{E}{2E_t} + 1 \right)^K \_ 1 \right] \qquad 2.4\text{-}13$$

The value of k can be calculated by matching the loudness of, first, a uniform masking noise (a noise with the same SPL in each critical band) to the loudness of a 1 kHz tone for various SPLs.

These values were then predicted from equation 2.4-13 by calculating the total loudness of each sound through the integral

$$N = \int_{Z=0}^{Z=24} N'dZ \qquad 2.4\text{-}14$$

for various values of k. The best fit was found for a value of K = 0.23

The value for $N'gr_o$ can be calculated by forcing the integral $\int N'dZ$ to be equal to 1 sone where N' is the specific loudness of a 1 KHz tone at 40 dB SPL.

After some smoothing at low levels the final equation is given by:

$$N' = 0.08 \left(\frac{E_t}{E_o} \right)^{0.23} \left[\left(\frac{1}{2} \frac{E}{E_t} + \frac{1}{2} \right)^{0.23} -1 \right] \qquad (11)$$

2.4-15

The equation can be seen in (figure 2.4-6) with both E_t and E expressed in dB as L_t and L_E. Note that the ordinate of fig [2.4-5] should not be labelled L_E but (L_E - a). a takes into account the frequency dependent attenuation introduced by the middle ear.

The model can be applied as it is to calculate the loudness of any complex sound provided E and E_t are known in each critical band. These can be derived from the masking pattern of the complex tone. This is not necessary though. The loudness can also be determined as follows: The SPL in each component critical band of the stimulus is measured. Each band is treated as an independent stimulus that gives rise to an excitation pattern like the one shown in fig [2.4-5]. the overlapping patterns thus obtained have a common upper envelope, which determines the excitation level; wherever two or more patterns overlap, only the highest excitation level is used.

The model can also predict partial masking if the slopes of the curves in figure [2.4-6] are adjusted as in figure [2.4-7]. In place of the excitation level L at threshold, the excitation level L_M of the masking stimulus is the parameter on the curves. Note that although the model is designed specifically for loudness calculation (summation) the basic elements (e.g. the critical band scale and the excitation patterns) can serve to model many other psychophysical phenomena. The model predicts values in good agreement with experimental results on loudness sumation [29] and partial masking [30].

The procedure has been standardized into a graphical form [9]. Figure [2.4-8] shows an example of factory noise whose octave band levels are shown. The area corresponds to the total loudness of the noise. The total area (the integral under the curve) is shown by the level of the horizontal line, two thirds down from the top.

2.4.3 The excitation pattern in other phenomena [31. 32]

Lets first review the elements of the model which will be relevant for the rest of the psychoacoustic phenomena to be examined. The first element is the critical bandwidth. To a first approximation, the critical band can be understood as a bandpass filter with infinitely steep slopes (rectangular filter). From the bandwidths, a scale can be derived such that one unit on this scale corresponds to the critical bandwidth. There is a nonlinear relationship between this scale and the frequency scale since the critical bandwidth increases with frequency. After the critical band scale is constructed one can do away with the rectangular shape filters and use the psychoacoustical excitation patterns instead. The advantage of the use of the excitation pattern on a critical band scale can be seen in figure [2.4-9]. The excitation for critical band wide noises of different center frequencies but equal SPL is shown. The curves are very similar and can be derived from each other by shifting them up or down the Z scale. The derivation of the excitation pattern for a narrow band noise, a broad band noise and 11 harmonics is shown in fig [2.4-10]. Another important fact is the nonlinearity of the upper slope of the excitation pattern as shown in figure (2.4-2). The äsymetry of the filter increases with level.

2.4.3.1 Difference Limens

The difference limens for both intensity and frequency can be explained with the following assumption:

"The ear is able to detect any change in a steady state sound if the excitation level L_E is changed anywhere along the critical band scale by the value $\Delta L_E >$ 1 dB."

The jnd for intensity jnd_I can be understood by taking into account that the tone evokes a broad excitation pattern including lower and upper accessory excitations. As was shown earlier the upper accessory excitation increases nonlinearly with intensity. This is the reason why the jnd_I is around 1 dB only for low loudness levels of 30 - 50 phon, while at higher levels the jnd_I drops down to 0.2 dB. In (figure 2.4-11a) the excitation for a tone with an intensity I and excitation level L_E is sown together with the excitation for a tone with an intensity I + ΔI and excitation $L_E+\Delta_E$. It can be seen that at higher levels where the upper accessory excitation is highly nonlinear a change in intensity has a larger effect on the excitation that at lower levels where the excitation behaves in a more linear fashion.

The jnd_f for frequency can be interpreted in a similar way (figure 2.4-11b). For changes in frequency the excitation pattern is shifted back and forth along the critical band scale. No nonlinearity is involved in this case. The largest change of the excitation level during the sifting occurs at the lower accessory excitation with the steep slope. Since the steepness of this slope is almost independent of both the level and the frequency of the exciting tone, the jnd_f should depend on frequency in the same way as the critical band scale z depends on frequency f. An average value for the slope of the lower accessory excitation is 27 dB/Bark. Assuming, again, a change ΔL_E = 1 dB for jnd, and since 1 unit on the z scale corresponds to one critical band the jnd_f is given by

$$\mathrm{jnd}_f = 1\ \mathrm{db}\ \frac{1\ \mathrm{Bark}}{27\ \mathrm{dB}} = \frac{1}{27}\ \Delta f_G \qquad 2.4\text{-}16$$

where Δf_G is the critical bandwidth at frequency f. The agreement of this prediction with experiment can be seen in (fig. 2.4-12).

2.4.3.2 Pitch and Phase Perception

The pitch scale (mel scale) is linearly related to the critical band rate scale (Bark).We have also seen that the steepest slope of the excitation pattern could be used to predict phase effects in monaural perception [10]. (Section 2.3.8)

2.4.3.3 Rougness

In [33] Vogel proposed a model or roughness summation similar to the loudness model of [26]. He found that the roughness of a single amplitude modulated tone decreased, when partially masked and therefore concluded that the fluctuation of the whole psychoacoustical excitation pattern is analysed for the sensation of roughness. He assumed that the excitation level fluctuates according to the degree of modulation at about the same amount at each place (neglecting the upper slope nonlinearity). Figure (2.4-13) shows the fluctuation of the excitation level pattern as a hatched area. He further assumed that the roughness r is composed of specific roughness r' which can be summed over the bark scale to give the roughness r. In equation form:

$$r(t) = \int_{z} r'(z,t)\,dz \qquad 2.4\text{-}17$$

note that both the roughness and specific roughness are functions of time.

2.4.4 Modelling Time Effects

Although the model in [26] was adequate to describe the loudness of steady sounds an extension to the model was necessary to encompass time dependent phenomena. First attempts to introduce a new part to the models to describe temporal effects in loudness and threshold can be found in [34].

There are various time constants involved in pschoacoustical phenomena [17] such as the time constant for loudness integration (around 200 ms) the time constant which limits the roughness sensation at high frequencies (around 15 ms) or the maximum time that is bridged over by forward and backward masking (around 2 ms). It is difficult to describe all the above phenomena with a single time constant and, therefore, in [34] a more complicated approach was taken. This model is shown in figure 2.4-14. The sound pressure p(t) is transferred to a filter bark which models the selectivity of the ear as in [26] by the excitation E as a function of tonalness z (critical band rate). Instead of using a very large number of outputs of the bank as would be the equivalent of the ear with its very many nerve fibers only 24 banks are used in regard to the 24 critical bands. Within or connected to the bank are rectifiers and square low transformers (note that this would give the short time envelope of the signal, see [35] p146 in combination with the following low pass filter). This gives the excitation $E_v(t)$ which is transferred to a RC network with a relatively short time constant T_1 (about 35 ms) integrating over a relatively short time (compared to loudness integration). It is then transformed into specific loudness and after passing it through a proportional differential transfer function it is finally passed through the last RC network with a time constant of T_2 = 10 ms. This final value is the specific loudness as a function of time N'(t)

The effects of the transformations can be seen in figure [2.4-15]. Note especially how the decay time is prolonged by 4 times when going from excitation through the power function to specific loudness. That this is so can be easily verified with an input $Ae^{-t/T}$ for the excitation E(t). After the nonlinear compression this becomes

$$N'(t) = A^{1/4} \cdot e^{-t/4T} \qquad 2.4\text{-}18$$

which obviously decays with a time constant 4 times than before. Note that the final value i.e. the time dependent specific loudness N'(t) only reaches the asymptote value for durations longer than 100 ms. It is important to note that the relevant output here is the peak of the N'(t) and not the integral of N'(t) over time. Temporal integration has already being performed and the result is the peak value of N'(t). This model appeared to describe loudness of variable sounds relatively well.

It now becomes interesting to know what the output of the filterbanks looks like when a stimulus with strong time structure is fed through. We are therefore concerned with the input to the time model described above. Some results can be found in [31]. These are very similar to the outputs of the filters shown in the section on pitch (figure 2.3-65]. They also help to explain some effects on timbre found by Plomp and Steeneken in [36].

A more comprehensive model but with essentially the same elements was presented in [37]. A good review of the time dependent phenomena that the model attempts to predict is also included. The model is as follows: The short time envelope of the incoming sound is obtained through analysis by a bank of critical band filters, a rectifier and a low pass filter. Specific loudnesses are obtained by amplitude compression and summed. The created value is then passed through two undescribed devices, a rectifier-non-linear-lowpass (NLLP) filter to simulate, postmasking effects and a "special" third order low pass filter whose constants are optimized in order to reproduce temporal partial masking. Unfortunately the reference relating to the special low pass filter is in German and no further mention of the NLLP is made in this paper. A full description of the NLLP is given in another paper though, in the form of an analogue circuit [38].

The model was used to predict the loudness of many sounds that vary both spectrally and temporally. The output of the device (loudness meter) to an input of tone bursts is shown in figure 2.4-16. Note that since the input can be confined to only one critical band this output represents what is analogous to an impulse response from a linear system. It can be seen that the results are not far different from those of fig [2.4-14] [34]. Again, the peak value represents the perceived loudness. This value correlates well to the perceived loudness of the different duration bursts. It also predicts well the loudness of AM tones, narrow band noise. temporally partially masked tone bursts, FM tones [39] and connected speech [40]. The latter is shown in (figure 2.4-17) together with the loudness of speech like-noise which elicits, according to Fastl [40], the same loudness. Again although the running speech shows a loudness pattern with a quite strong temporal pattern, the hearing system seems to perceive speech, in such a way that the loudest parts of the spoken sentences are responsible for loudness. Once again the loudness is determined from the peaks of the loudness meter output.

An extended model to the above was presented in [38]. This model is in a sense much more interesting from the one in [37] since it is constructed for use in automatic speech recognition and obtains values for loudness, roughness, pitch, signal duration and timbre through a front-end model very similar to that in [37]. Again review of the different sensations the model attempts to predict is given in the paper. The entire model is given in figure 2.4-18, whilst the front end processor is shown in figure (2.4-19). The main difference form the model in [37] is the shift of the non-linear low pass filter (here denoted as NLD-nonlinear device) before the point of loudness summation and thus to each channel individually. Also additional stages are included to produce roughness, and sharpness as well as loudness. The

NLD is shown in figure [2.4-20]. Since the circuit contains elements which behave in a nonlinear fashion it cannot be analysed through z-transforms. It can be digitally simulated by splitting it into three separate phases, either $e_1(t) = e_2(t)$, or, if $e_1(t)<e_2(t)$ then into $e_3(t)<e_2(t)$ or $e_3(t) = e_2(t)$. $e_3(t)$ is the voltage at the common point of the 1μF capacitor and the 20 K resistor. Each phases's new sample $e_2(t_1)$ can be calculated from each past sample $e_2(t_0)$ through difference equations.

The outputs of the front end processor from different input stimuli are shown in figure 2.4-21 and 2.4-22. Limited success is reported from use of the model for speech recognition.

To end the review of models by Zwicker, we mention a model to predict masking period patterns [41]. This is not unlike the previous ones [37, 38] but more care is taken on phase conditions and, some new elements such as an interfering noise source and the nonlinearity of the upper slopes of the excitation patterns are taken into account.

2.4.5 Schroeder's et al model

In [42] Schroeder offered some algebraic formulas for the various transformations given graphically in Zwicker and Scharf's model [26]. These concerned a transformation from the frequency scale to the tonalness (Bark) scale and an expression for the excitation pattern of subcritical stimuli. These expressions were later used to provide an algebraic model for predicting the loudness of quantization noise in speech signals. This model was based largely on the model in [26] but, also, some new results were incorporated from [43] concerning the masking ability of tone signals upon noise. The model was then applied in a predictive coder [41] (figure 2.4-23) to reduce the perceptual impact of quantization noise.

The time varying feedback filter F in figure 2.4-23 was iteratively optimized for each frame to minimize the objective noise loudness as given by the model. The model is as follows:

The power spectra of the speech signal $\tilde{S}(f)$ and the Noise $\tilde{N}(f)$ are computed over time windows of approximately 20 ms duration.

The power spectra obtained are then transformed into "critical band densities". Note that $\tilde{S}(f)$ is the power density function dJ/df where J is the total power in the speech signal given by

$$J = \int_F \tilde{S}(f)df \qquad 2.4\text{-}19$$

where F is the frequency region where $\tilde{S}(f)$ has a significant contribution to the integral. If now, we wish to express the power density over some other scale related to the frequency scale, and in this case over the critical band rate scale, x then

$$S(X) = \frac{dJ}{dx} = \frac{dJ}{df}\,\frac{df}{dx}$$

$$= \tilde{S}(f)\,\frac{df}{dx} \qquad 2.4\text{-}20$$

the relationship between f and x is given by

$$f = 650 \sinh\left(\frac{x}{7}\right) \qquad 2.4\text{-}21$$

and, if the final expression for S(X) is to contain only x explicitly

$$S(X) = \tilde{S}(f(X)) \quad \frac{df}{dx} \qquad 2.4\text{-}22$$

similarly for the noise

$$N(x) = \tilde{N}(f(x)) \quad \frac{df}{dx} \qquad 2.4\text{-}23$$

Next, excitation patterns are computed by controlling the critical band densities with a "spreading" function B(X) which describes the excitation pattern of subcritical stimuli

$$E(X) = S(X) * B(X) \qquad 2.4\text{-}24$$

An expression for B(X) is

$$10\log_{10} B(X) = 15.81 + 7.5\ (X + 0.474)$$

$$- 17.5\ (1 + (X + 0.474)^2)^{½}\ \text{dB} \qquad 2.4\text{-}25$$

and the noise excitation pattern is derived in a similar way:

$$Q(X) = N(X) * B(X) \qquad 2.4\text{-}26$$

Then, the loudness of the speech signal is computed from

$$L_s = C \int_X [E(x)]^{0.25}\, dx \quad \text{sone} \qquad 2.4\text{-}27$$

where again X is the range of X for which the integral yields a finite value.

The constant C can be chosen so that the units of loudness can be calculated in sones.

At the threshold of hearing, the loudness is zero. A formula to describe the effects near threshold is given as

$$L_s = C \int_X \mathrm{Max}\ \{[E(x) - \theta(X)]^{0.25};\ 0\}\ dx \qquad 2.4\text{-}28$$

where $\theta(X)$ is the threshold of hearing at each value of X.

The loudness of noise is reduced due to partial masking. A proposed formula for its calculation was given as:

$$L_N = C \int_X \left(\frac{Q(X)}{1 + [E(X)/Q(X)]^p}\right)^{0.25} dx \qquad 2.4\text{-}29$$

where $p \approx 2$

Note that the range X for the noise loudness need not be the same as the range for the speech loudness.

To include threshold effects for the case of the speech masker, the masked threshold is calculated from

$$M(X) = W(X)\ E(X) \qquad 2.4\text{-}30$$

where W(X) is a sensitivity function.

An expression for W(X) from complete masking experiments carried out by the authors is given by

$$10 \log_{10} W(X) \approx (15.5 + X)\ \mathrm{dB} \qquad 2.4\text{-}31$$

The behaviour near threshold (either masked or physiological) is given by

$$L_N = C \int_X \frac{\text{Max}[Q - \text{Max}(M;\theta);0]^{0.25}}{(1 + E/Q)^P} dX \qquad 2.4\text{-}32$$

The objective measure is then given by

$$D = L_N/L_S \qquad 2.4\text{-}33$$

as a measure of speech degradation. The division of L_N by L_S is rather interesting since the effect of partial (or complete) masking due to the speech signal has already been taken into account. It implies that noise of the same loudness is more objectionable (hence louder!?) when the speech loudness is lower than otherwise, which suggests that noise loudness is not perhaps the relevant criterion.

The model does not take into account temporal effects, of which temporal summation of loudness is probably the most relevant. The loudness is also inherently time varying since the parameters are computed periodically. It is difficult to relate to the expression "loudness of speech" given by the following equation

$$L_s = C\int_X [E(X)]^{0.25} \, dx \qquad 2.4\text{-}34$$

when the loudness of speech is determined from the high energy segments over a period comparable to the order of seconds and not 20 msec [40,37].

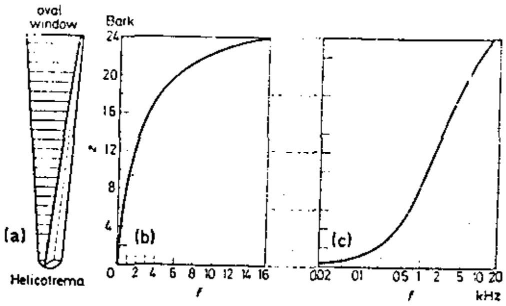
The noise loudness would be a relevant criterion only when the "noise" is perceived as background uncorrelated noise (i.e. hiss) typical to PCM or differential coders of relatively high bit-rate. When this noise manifests itself in other forms such as e.g. roughness then we have seen that other models would be more applicable. Worse still, in the more sophisticated coders, the noise is not really "heard" but, instead,alters the quality of the speech signal in such a way that sounds different from the original, but cannot be described as "noisy". In such cases, to proceed to calculate the loudness of the noise would be meaningless.

Other models using psychoacoustic knowledge also appeared in the literature e.g. [44], [45], but at large, they are mainly concerned with speech envelope quantization rather than the whole signal.

Table 2.4-T1 [15]

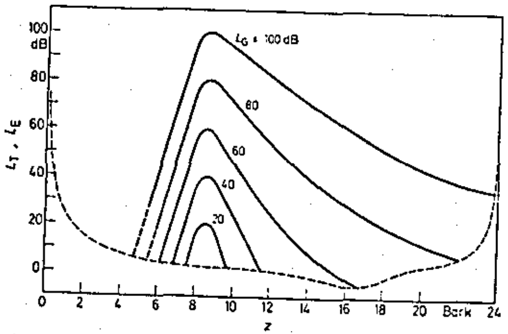
Values of critical band rate and critical bandwidth as functions of frequency.

Critical band rate, Bark	Frequency Hz	Critical bandwidth Hz	Center frequency Hz
0	0		
		100	50
1	100		
		100	150
2	200		
		100	250
3	300		
		100	350
4	400		
		110	450
5	510		
		120	570
6	630		
		140	700
7	770		
		150	840
8	920		
		160	1000
9	1080		
		190	1170
10	1270		
		210	1370
11	1480		
		240	1600
12	1720		
		280	1850
13	2000		
		320	2150
14	2320		
		380	2500
15	2700		
		450	2900
16	3150		
		550	3400
17	3700		
		700	4000
18	4400		
		900	4800
19	5300		
		1100	5800
20	6400		
		1300	7000
21	7700		
		1800	8500
22	9500		
		2500	10 500
23	12 000		
		3500	13 500
24	15 500		



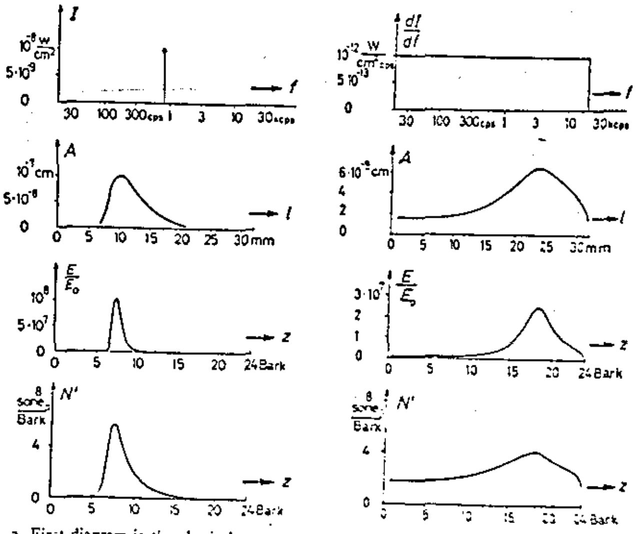
Critical band rate, z, as a function of the frequency, f, given on linear (b) and logarithmic scale (c). The drawing (a) indicates the relation to the locations along the basilar membrane from the oval window to the helicotrema

Figure 2.4-1 [32]



Excitation pattern (excitation level, L_E, as a function of the critical band rate, z) of a critical band wide noise at different levels, L_G, with a constant center frequency of 1 kHz corresponding to 8.5 Bark (solid). Level, L_T, of a pure tone at threshold in quiet (dashed)

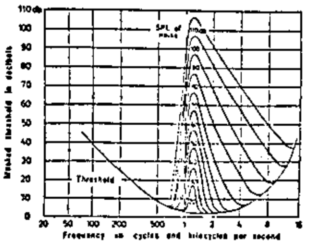
Figure 2.4-2 [32]



a. First diagram is the physical spectrum of an 800-cps tone. (Intensity I plotted against frequency f. Next is plotted the Amplitude A of the displacement caused by the tone on the basilar membrane as a function of Distance l from the helicotrema. Third is shown the Excitation E at the Organ of Corti as a function of tonalness z. [The symbol E_0 is an arbitrary reference.] Last is shown the Specific Loudness N' as a function of Tonalness z. The area under the loudness pattern corresponds to the total loudness of the tone.)

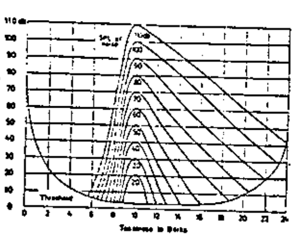
b. The spectrum of a white noise followed by the displacement, excitation, and loudness patterns produced by the noise.

Figure 2.4-3 [26] Sequential formation of a loudness pattern.



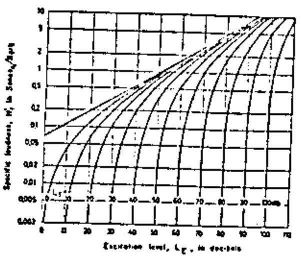
Masking effect of a narrow band of noise with a center frequency of 1,200 cps. The parameter is the effective SPL of the noise. On the ordinate the masked threshold for a pure tone is plotted for each level of noise as a function of frequency.

Figure 2.4-4 [26]



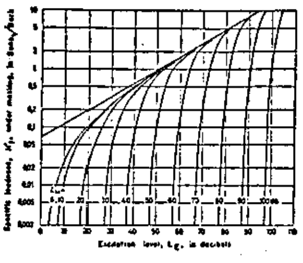
Excitation patterns calculated from the masking produced by a narrow band of noise centered at 1,200 cps. The level of the effective SPL of the band of noise is the parameter on the curves.

Figure 2.4-5 [26]



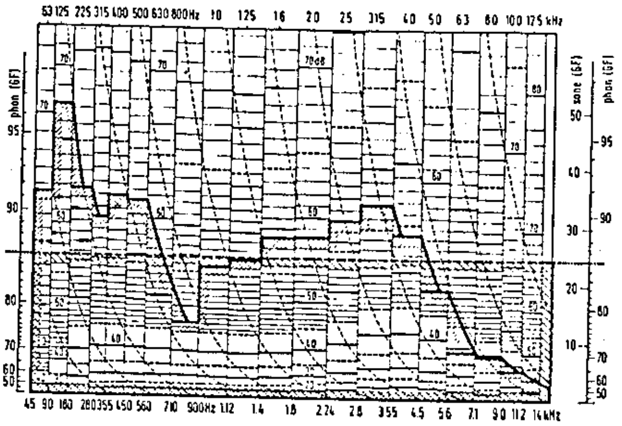
Specific Loudness N' in sone$_G$/Bark as a function of the Excitation Level L_E. The exponent is $k = \frac{1}{4}$. The parameter on the curves is the Excitation Level L_t of the test tone at threshold.

Figure 2.4-6 [26]



Specific Loudness N'_t of partially masked pure tones as a function of the excitation level of the tone. The parameter on the curves is the Excitation Level L_M of the masking sound.

Figure 2.4-7 [26]



Calculation of the loudness of factory noise from its third octave band levels using a graphical method based on the described model

Figure 2.4-8 [32]

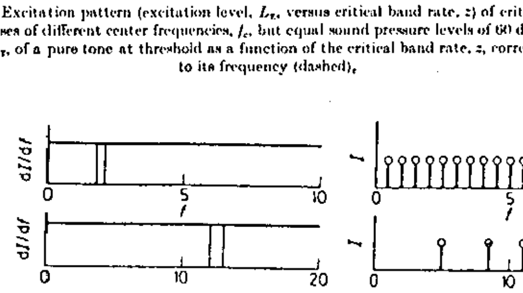


Figure 2.4-9 [32]

Excitation pattern (excitation level, L_E, versus critical band rate, z) of critical band wide noises of different center frequencies, f_c, but equal sound pressure levels of 60 dB (solid). Level, L_T, of a pure tone at threshold as a function of the critical band rate, z, corresponding to its frequency (dashed).

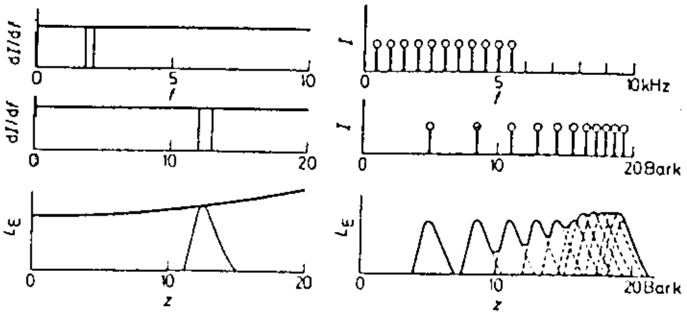
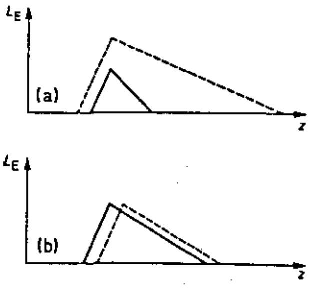


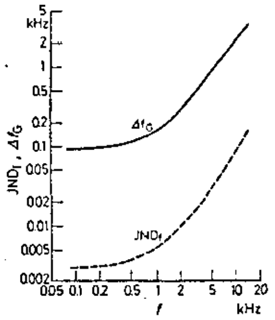
Figure 2.4-10 [32]

Derivation of the excitation pattern for broad band noise and narrow band noise (left) as well as for a sound consisting of eleven harmonics (right)



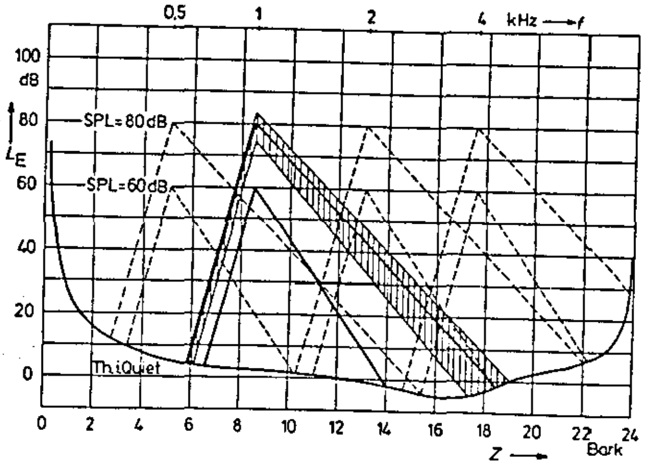
Schematic drawings of the change of the excitation pattern, L_E (z), produced by a change in level (a) and produced by a change in frequency (b)

Figure 2.4-11 [32]



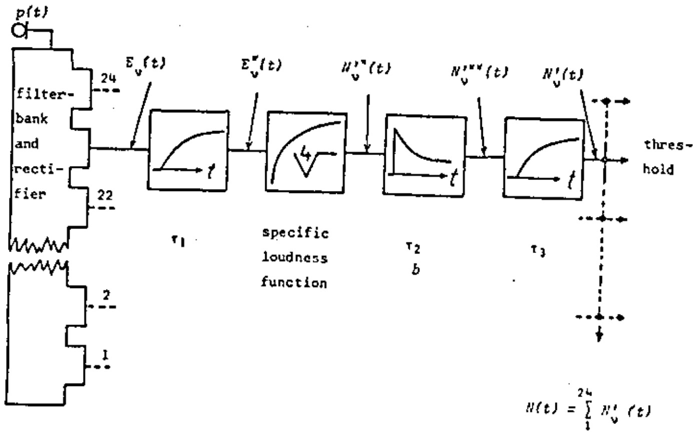
Band width, Δf_G, of the critical band (solid) and just-noticeable difference, JND_f, for frequency (dashed) of pure tones as a function of the frequency, f

Figure 2.4-12 [32]



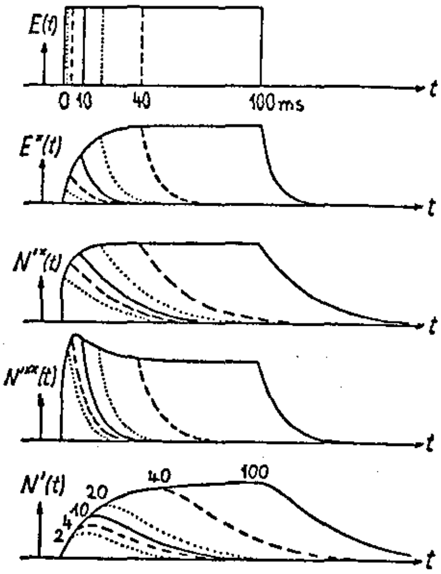
Excitation level - critical band rate - pattern of narrow band sounds of various frequencies and various SPL's. The hatched range characterizes the fluctuation range of a modulated tone (m=0,5).

Figure 2.4-13 [33]



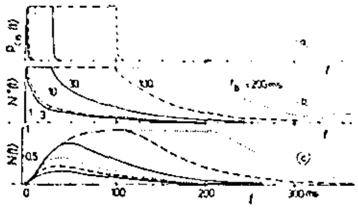
Function model.

Figure 2.4-14 [34]



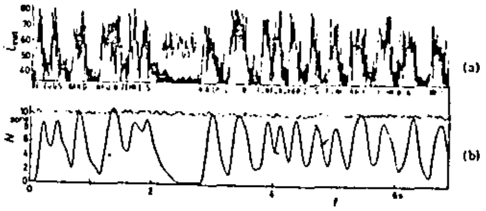
Transformation of $E(t)$ into $N(t)$ illustrated on impulses of different duration.

Figure 2.4-15 [34]



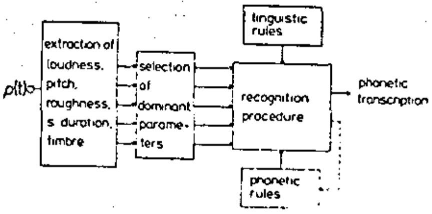
Loudness development of single tone bursts with durations $T_B = 1$, 3, 10, 30, 100, and 200 ms. The critical-band sound pressure p_{CB}, the intermediate value N^*, and the loudness N are shown as a function of time t in (a), (b), and (c), respectively.

Figure 2.4-16 [37]



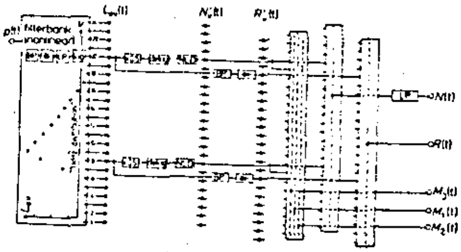
Instantaneous sound pressure level $L_{i\,ms}$ (a) and loudness-time function (b) produced by the loudness meter for running speech (solid lines) and for speechlike noise (dotted lines) which are, according to Fastl (1976), equally loud.

Figure 2.4-17 [37]


p(t)
extraction of loudness, pitch, roughness, s duration, timbre
selection of dominant parameters
linguistic rules
recognition procedure
phonetic transcription
phonetic rules

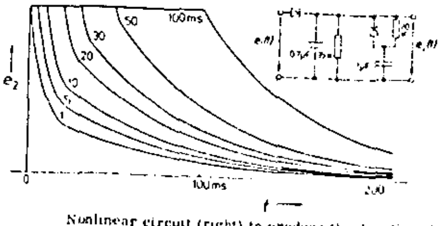
Schematic outline of the whole speech recognition system. $p(t)$ = signal input (sound pressure).

Figure 2.4-18 [38]



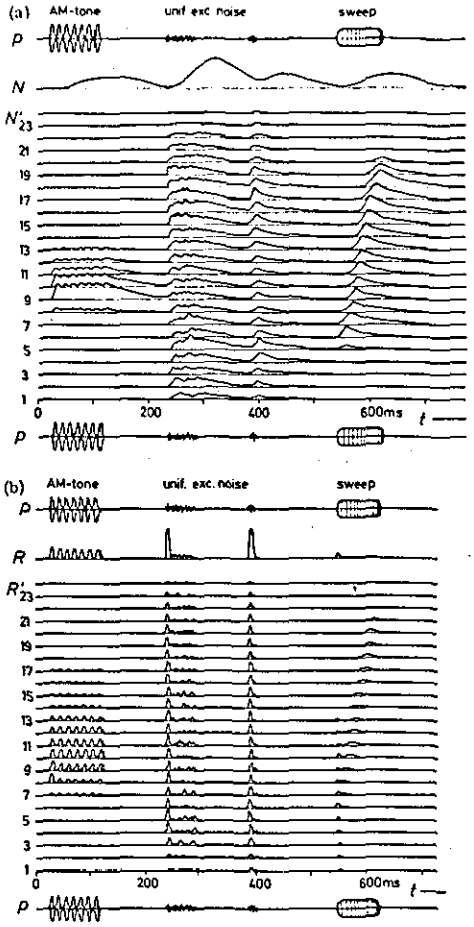
Block diagram of the preprocessing system which has been realized in analog hardware. $L_e(t)$: excitation level. $N'(t)$: specific loudness. $R'(t)$: specific roughness. $N(t)$: total loudness. $R(t)$: total roughness. M_0, M_1, M_2: momenta of specific loudness. BP: bandpass-filter. LP: low-pass filter. NLD: nonlinear circuit.

Figure 2.4-19 [38]



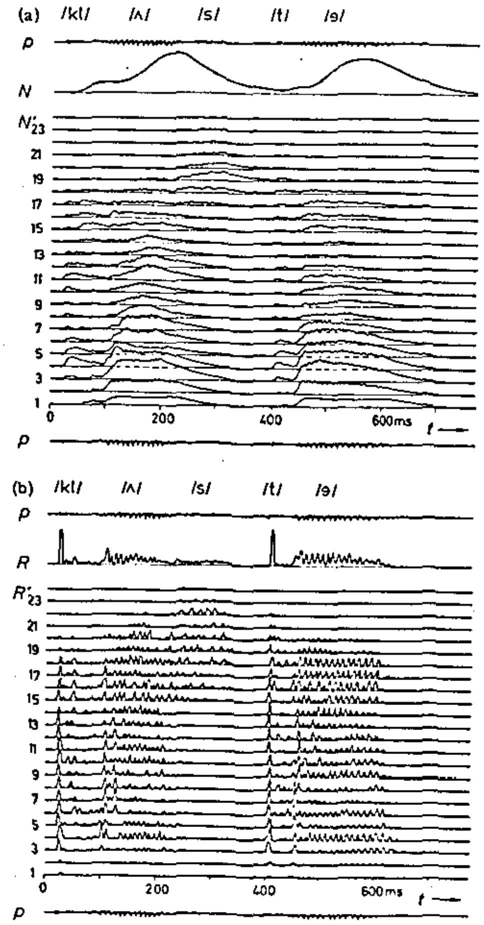
Nonlinear circuit (right) to produce the duration-dependent decay as shown on the left. Parameter is the duration of the DC-impulse at the input.

Figure 2.4-20 [38]



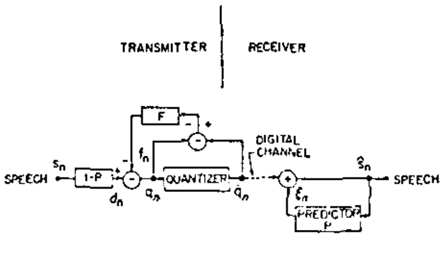
(a) Signal processing by the hardware preprocessing system. Input signal, $p(t)$, total loudness, $N(t)$, and specific loudness, $N'(t)$, as produced by an AM tone, impulses of uniform exciting noise, and a frequency-sweep tone, respectively (see text). (b) As (a), with total roughness, $R(t)$, and specific roughness, $R'(t)$, respectively.

Figure 2.4-21 [38]



Signal processing of the hardware pre-processing system. Processing of the utterance "cluster."

Figure 2.4-22 [3S]



Block diagram of a generalized predictive coder.

Figure 2.4-23 [41]

1. J. Zwislocki "Theory of Temporal Auditory Summation" J. Acoust. Soc. Am. Vol. 32, No. 8 August 1960 pp1046-1059.

2. J. Zwislocki "Temporal Summation of Loudness: An Analysis" J. Acoustical Soc. Am. Vol. 46, No. 2 part 2, 1969 pp431-441.

3. N.R. French and J.C. Steinberg "Factors Governing the Intelligibility of Speech Sounds" Vol. 19, No. 1, Jan. 1947 pp90-119.

4. S. Saito, S. Watanabe "Normalised Representation of Noise-Band Masking and its application to the prediction of Speech Intelligibility" J. Acoust. Soc. Am. Vol. 33, No. 8, Aug. 1961 pp1013-1021.

5. K.D. Kryter "Methods for the calculation and use of the Articulatin Index" Vol. 34, No. 11, Nov. 1962, J. Acolust. Soc. Am. pp1689-1697.

6. K.D. Kryter "Validation of the Articulation Index" J. Acoust. Soc. Am. Vol. 34 No. 11, 1962 pp1698-1702.

7. S.S. Stevens "Procedure for Calculating Loudness: Mark VI" J. Acoust. Soc.Am. Vol. 33, No. 11 1961, pp1577-1585.

8. International Standards REcommendation R131 Sept. 1959.

9. International Standards Recommendation R532 Sept. 1966.

10. J.L. Goldstein "Auditory Spectral filtering and Monaural Phase Perception" J. Acoust. Soc. Am. Vol. 41, No.(2) 1967. pp458-479.

11. H. Duifhuis "Consequences of peripheral frequency selectivity for nonsimultaneous masking" J. Acoust. Soc. Am. Vol. 54, No. 6, 1973 pp1471-1488.

12. H. Duifhuis "A crude Quantitative theory of backward masking" In: Facts and Models in Hearing E. Zwicker, E. Terhardt Eds. Springer-Verlag 1974 pp275-284.

13. E. deBoer "On the "Residue" and Auditory Pitch Perception" In: Handbook of Sensory Physiologyl Vol. v/3, Springer Verlag Berln, Heidelberg, New York, 1976 pp479-583.

14. B. Scharf "Critical Bands" In: Foundations of Modern Auditory Theory" J.V. Tobias (Ed.) Vol. I, 1970 pp159-202.

15. E. Zwicker and E. Terhardt "Analytical Expession for critical-band rate and critical bandwidth as a function of frequency" J. Acoust. Soc. Am. 68, 1980, (Letter to the Editor) pp1523-1525

16. M.R. Schroeder, B.S. Atal, J.L. Hall "Optimizing digital speech coders by exploiting masking properties of the human ear" J. Acoust. Soc. Am. 66(6) Dec. 1979 pp1647-1652.

17. E. Zwicker "Temporal effects in Psychoacoustical Excitation" in "Basic Mechanisms in Hearing" A.R. Moller (ed.) Academic Press 1973 pp809-827.

18. J.A. Swets, D.M. Green, W.P. Tanner, "On the width of Critical Bands" J. Acoust. Soc. Am. Vol. 34 No. 1, 1962 pp108-113.

19. R.D. Patterson "Auditory filter shape" J. Acoust. Soc. Am. Vol. 55, No. 4, 1974 pp802-809.

20. R.D. Patterson "Auditory filter shapes derived with noise stimuli" J. Acoust. Soc. Am. Vol. 59, No. 3, 1976 pp640-654.

21. R.D. Patterson and Ian Nimmo-Smith "Off-frequency listening and auditory filter asymmetry" J. Acoust. Soc. Am. 67(1) Jan 1980, pp229-245.

22. B.C.J. Moore and B.R. Glasberg "Auditory filter shapes derived in simultaneous and forward masking" J. Acoust. Soc. Am. 70(4) Oct. 1981, pp1003-1014.

23. R.D. Patterson, Ian Nimmo-Smith, D.L. Weber, and R. Milroy "The deterioration of hearing with age: Frequency selectivity the ciritical ratio, the audiogram and speech threshold" J. Acoust. Soc. Am. 72(6) 1982 pp1788-1803.

24. B.C.J. Moore, B.R. Glasberg "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns" J. Acoust. Soc. Am. 74(3) 1983 pp750-753.

25. B.R. Glasberg and B.C.J. Moore "Comparison of auditory filer shapes derived with three different maskers" J. Acoust. Soc. Am. Vol. 75(2) Feb. 1984 pp536-544

26. E. Zwicker, B. Scharf "A model of Loudness Summation" Psychological Review, 1965, Vol. 72, No. 1, 3-26.

27. D.D. Greenwood "Auditory Masking and the Critical Band" Vol. 33, No. 4, 1961 pp484-502.

28. Steven, S.S. "On the psychophysical law" Psychophysical Review, 1957, 64, 153-181.

29. Zwicker, G. Flottorp, G. and Stevens S.S. "Critical bandwidth in loudness summation" J. Acoust. Soc. Am. 1957, 29 pp548-557.

30. B. Scharf "Partial Masking" Acustica 1964, 14, pp16-23.

31. E. Zwicker "Masking and psychological excitation as consequencies of the ear's frequency analysis" in: Frequency analysis and periodicity detection in hearing" R. Plomp and .F. Smoorenburg (eds.) A.W. Sijthoff - Leiden 1970 pp376-396.

32. E. Zwicker "Scaling" In: Handbook of Sensory physiology V/2 Springer-Verlag 1975 pp401-448.

33. A. Vogel "Roughness and its relation to the time-pattern of psychoacoustical excitation" In: Facts and Models of Hearing, E. Zwicker and E. Terhardt (eds.) Springer-Verlag pp241-250.

34. E. Zwicker "A model describing temporal effects in loudness and threshold" The 6th International Congress on acoustics Tokyo, Japan, 1968 A-3-4, A-21-A24.

35. J.L.Flanagan "Speech Analysis Synthesis and Perception" Springer-Verlag 1972.

36. Plomp, R. and Steeneken, H.J.M. (1969) "Effect of phase on the timbre of complex tones" J. Acoust. Soc.Am. 46, 409-421.

37. E. Zwicker, "Procedure for calculating loudness for temporlly variable sounds" J. Acoust. Soc. Am. Vol. 62, No. 3, Sept. 1977.

38. E. Zwicker, E. Terhardt, E. Paulus, "Automatic Speech Recognition using psychoacoustic models" J. Acoust Soc. Am. 65(2) 1979 pp487-498.

39. E. Zwicker "Loudness and excitation patterns of strongly frequency modulated tones" in: Sensation and Measurement, H.R. Moskowitcz, B. Scharf, J.C. Stevens (eds.) D. Reidel, Dordrecht/Boston pp325-335 1974.

40. Fastl, H. "Loudness of running speech" J. Audiol. Technique 16, pp2-13 1977.

41. B.S. Atal, M.R. Schroeder "Optimizing predictive coders for minimum Audible Noise" Inter. Conf. Acoust. Speech. Signal proc. IEEE 1979, 453-455.

42. M.R. Schroeder "Recognition of complex acoustic signals" Life Sciences Research Report 55, T.H. Bullock, Ed. pp323-328, 1977.

43. R.P. Hellman "Asymmetry of masking between noise and tone" perception and Psychophysics 1972 Vol. 11(3) pp241-246.

44. O.Ghitza and J.L,Goldstein "Scalar LPC Quantization based on Formant Jnds" IEEE trans on Acoust. Speech Signal Proc. Vol. ASSP-34, No. 4, 1986, pp697-708.

45. D.B. Paul "An 800bps adaptive vector quantization vocoder using a perceptual distance measure" Int. Conf. Acoust. Speech Signal Proc. IEEE pp72-76, Boston 1983.

CHAPTER 3

SPEECH AND SPEECH CODING

3.1 SPEECH: PRODUCTION AND PHONETICS

3.1.1 Acoustical Speech Production

The acoustical speech waveform is an acoustic pressure wave which originates from voluntary physiological movements of the structures shown in figure (3.1-1a). Air is expelled from the lungs into the trachea and then forced between the vocal folds (cords). During the generation of voiced sounds, air flow from the lungs is modulated by the vocal cord vibration resulting in a quasi-periodic pulse-like excitation. As a periodic signal, voiced speech has spectra consisting of harmonics of the fundamental frequency of the vocal fold vibration; This frequency often abbreviated FO is the physical aspect of the speech signal corresponding to perceived pitch. The harmonics are energy concentrations at multiples of FO. The average values for FO are around 130 Hz for males and 220 Hz for females.

Unvoiced sounds are generated by voluntarily holding the vocal cords open, forcing air past them, and then using the articulators to create a constriction.

The air flow from the lungs becomes turbulent as the air passes through the constriction resulting in a noise-like aperiodic excitation.

Another mode of excitation occurs when air-flow builds up pressure behind a point of total closure in the vocal tract. The rapid release of this pressure, by removing the constriction, causes a transient excitation [2]. The opening between the vocal cords is defined as the glottis. The vocal tract is a non-uniform acoustic tube which extends from the glottis to the lips and varies in shape as a function of time. The major anatomical components causing this time varying change are the (articulators) lips, jaw, tongue and velum. During the generation of the non-nasal sounds the velum closes off the vocal react from the nasal cavity. The nasal cavity constitutes an additional acoustic tube for sound transmission used in the generation of nasal sounds. As sound generated as discussed above propagates down this tube, the frequency spectrum is shaped by the frequency selectivity of the tube. This effect is very similar to the resonance effects observed with wind instruments. The resonance frequencies of the vocal tract are called formants. The formant frequencies depend upon the shape and dimensions of the vocal tract. Each shape is characterized by a set of formant frequencies. Different sounds are formed by varying the shape of the vocal tract. Thus, the spectral properties of the speech signal vary with time as the vocal tract shape varies.

In the average male, the total length of the vocal tract is about 17 cm. The cross-sectional area, determined by the articulators, tongue, lips, jaw and velum varies from zero to 20 cm^2. Although the formant frequencies are primarily related to the shape of the vocal tract, there is some frequency shift due to losses. The bandwidths of the formants are determined from these losses which include losses due to the softness of the vocal tract walls, viscous friction of the air, thermal losses and radiation loss. The formants are abbreviated Fi where F1 is the formant with the lowest center frequency.

3.1.2 Phonemes

Most languages can be described in terms of a set of distinctive sounds or phonemes. In the english language there are about 40 phonemes. A table of phonemes is given in Table (3.1-T1). As can be seen from the table the phonemes are broken into various classes. The manner of articulation is concerned with airflow: for vowels and diphthongs the airflow meets no constriction narrow enough to cause turbulent flow (frication). Glides (semivowels) are similar to vowels but employ narrow vocal tract constrictions that may cause frication. Liquids too are similar to vowels but use the tongue or an obstruction in the oral tract causing air to deflect around the tip. During nasal sounds the velum is lowered and its position allows airflow out of the nostrils.

All of the above phonemes employ voicing and excite the vocal tract solely at the glottis; these continuous, intense and periodic phonemes are also known as sonorants. The remaining obstinent phonemes are weak and aperiodic and are primarily excited at their major vocal tract constriction. Stops involve a rapid closure of a vocal tract obstruction, a pressure build up and a sudden release with a rush of air that creates a brief (e.g. 10 ms) acoustic burst. Frigatives employ a narrow constriction. The vocal tract is excited by a steady flow which becomes turbulent in the region of the constriction.

The phenomes are also characterized by their place of articulation. This relates the phonemes to the point in the vocal tract of narrowest constriction. These places are shown in figure (3.1-1b). The relation of the phonemes to the place of articulation can be seen in figure (3.1-2).

3.1.3 Acoustic Phonetics

The phonemes are also distinguished from their waveform and spectral properties. The waveforms can be viewed directly whereas there are several ways to present the spectral content of each phoneme. One way is to use the formant frequencies [3]. This method is particularly useful for the study of vowel sounds. Another popular way is through the use of voice spectrograms. These display time and frequency on the horizontal and vertical axes whereas amplitude (logarithmically compressed) is related to the darkness of the display. Formant frequencies appear as dark horizontal bands.

The first two formants are usually sufficient to distinguish amongst vowels. This can be seen from figure (3.1-3). A diphthong is a gliding monosyllabic speech item that starts at or near the articulatory position for one vowel and moves to or toward the position for another. This can be seen on the F1-F2 plane in figure (3.1-4). Spectrograms for vowels are shown in figure (3.1-5) whereas for consonants and stops and nasals are shown in figures (3.1-6) and (3.1-7). Note that although for vowels we have only resonance frequencies, for nasals the mouth serves as a resonant cavity that traps acoustic energy at certain natural frequencies. As far as the radiated sound is concerned, these resonant frequencies appear as antiresonances, or zeros of sound transmission. Waveforms for vowels and consonants are shown in figures (3.1-8) and (3.1-9).

3.1.4 Modelling the speech production process

Although the shape of the vocal tract changes continuousy, it can be considered relatively fixed or quasi-stationary over short periods of time (tens of milliseconds). This is short enough to account for the duration of stops and other phonemes which involve motion of the vocal tract. For vowels this duration can be extended from a range of 50 to 400 ms. Over this short time interval of a few tens of milliseconds the vocal tract can be modeled as a filter which appears as time varying over longer durations of time. The resonance frequencies are primarily determined by the way the cross-sectional area varies along the vocal tract. The dependence of cross sectional area upon distance along the vocal tract is called the area function of the vocal tract. This is determined from the position of the tongue, jaw, lips and velum: For each sound there is a positioning for each of the vocal tract articulators. Figure (3.1-10) shows the general block diagram that is representative of numerous models of speech production. The common element in these models is that the excitation features are separated from the vocal tract and radiation features. The vocal tract and radiation effects are accounted for by the time varying linear system. Its purpose is to model the resonance effects of the vocal tract. A widely used model is based upon the assumption that the vocal tract can be represented as a concentration of lossless acoustic tubes, figure (3.1-11). The constant cross sectional areas A_K of the tubes are chosen so as to approximate the area function of the vocal tract. The approximation improves as the number of sections increases. This model, of course ignores losses due to friction, heat, conduction and vocal react wall vibration. Closely related to the A_K is the reflection coefficient

$$r_K = \frac{A_{K+1} - A_K}{A_{K+1} + A_K} \qquad 3.1-1$$

Since A_K is positive then $-1< r_K <1$. The reflection coefficient r_K gives the relative amount of the travelling wave (travelling from the glottis to the lips) that is reflected back at the junction between the tubes with areas A_{K+1} and A_K. The reflection coefficients at each junction are an alternative form of representing the vocal tract model. It can be shown [2,5] that the relationship between input and output of the model can be represented by a transfer function V(Z) in the form

$$V(Z) = \frac{G}{1 - \sum_{K=1}^{N} a_K Z^{-K}} \qquad 3.1\text{-}2$$

where G and $\{a_K\}$ depend upon the area function.

The resonances (formants) of speech correspond to the poles of the transfer function V(Z). This all pole model is a very good representation of vocal tract effects for vowels but nasal and fricatives require both poles and zeros. The effect of a zero can also be achieved by including more poles [6] since

$$1 - aZ^{-1} = \frac{1}{\sum_{n=o}^{\infty} a^n Z^{-n}} \qquad 3.1\text{-}3$$

Since the coefficients of the denominator of V(Z) are real the roots will be either real or occur in complex conjugate pairs. Therefore, there will be at most N/2 resonances (formants). A complex resonant frequency is given by

$$S_K,\ S_K{*} = -\sigma \pm j2\pi F_K \qquad 3.1\text{-}4$$

in the S plane or

$$Z_K,\ Z_K{*} = e^{-\sigma_k T} . e^{\pm j2\pi F_k T} \qquad 3.1\text{-}5$$

$$= e^{-\sigma_K T} \cos (2\pi F_K T) \pm j e^{-\sigma_K T} \sin (2\pi F_K T) \qquad 3.1\text{-}6$$

in the Z-plane (figure 3.1-12).

The bandwidth of the resonance is approximately $2\sigma_K$ and the centre frequency is $2\pi F_K$. In the Z-plane, the radious from the origin to the pole determines the bandwidth i.e.

$$|Z_K| = e^{-\sigma_K T} \qquad 3.1\text{-}7$$

and the Z-plane angle determines the centre frequency

$$\theta_K = 2\Pi F_K T \qquad 3.1\text{-}8$$

For stability $\sigma_K > 0$ or alternatively $|Z_K| < 1$ 3.1-9

As long as the areas A_K are positive then all the poles will be inside the unit circle [5,6]. and the model will be stable.

The first attempt at directly computing an acoustic tube model of the vocal tract from the speech waveform is due to Atal [7]. He demonstrated that formant frequencies and bandwidths are sufficient to uniquely determine the area of an acoustic tube having a specified number of sections. He also demonstrated that a transfer function with N poles is always realizable as the transfer function of an acoustic tube consisting of N cylindrical sections of equal length. Wakita [8] showed that the same acoustic tube model is equivalently represented from the inverse filter A(Z) obtained by linear prediction of the acoustical speech waveform. He also demonstrated the important experimental result that if the speech is properly preemphasized and if boundary conditions of the acoustic tube are properly chosen, then very reasonable vocal tract shapes can be directly estimated using the autocorrelation method of linear prediction. The effects of glottal waves and radiation can be reasonably well estimated by fixed preemphasis of the speech since their characteristics vary relatively slowly in the frequency domain. An analysis example of the acoustic waveform, the filter's V(Z) spectral response and the area functions are given in figure 3.1-13.

3.2 SPEECH CODING

3.2.1 Introduction

Digital transmission and storage of speech signals has become dominant over and will very soon replace, most analog systems. This is entirely due to recent advances in VLSI (and other) technology which made digital methods cost effective in comparison to their analog counterparts. The transmission bit rate is a crucial factor in evaluating the practicality of different coding schemes: The bandwidth of a transmission channel limits the number of signals that can be carried simultaneously. The lower the bit rate for a speech signal the higher the number of signals that can be carried simultaneously. Similarly for voice storage (e.g. electronic mail) lower bit rates reduce the computer memory needed to store the speech signals. The cost of digital transmission and storage systems such as optical fibers and computer memory has seen a dramatic reduction is recent years. Also the suitability of digitally encoded signals for processing by digital single-chip computers has brought about a large reduction in the bit rate required with little loss of quality. The wider availability and reduction in price of these signal processing chips in recent years has stimulated research into new, more efficient algorithms for speech coding.

The cost effectiveness of digital processing and transmission does not make the digital coding of speech signals necessarily desirable, it merely makes them feasible. There are several other factors which determine the superiority of digital over analog transmission. It is relatively easy to apply encryption techniques to digital signals and provide privacy, which is one of the reasons why digital coding of speech was popular for military applications at a time when it was not cost-effective for commercial use. Digital encoding enables transmission of information over long distances to be achieved without degredation of the speech quality. Analog transmission channels always distort audio signals to a certain extent but digital communication links can regenerate i.e. retime and reshape the signals at repeaters placed along the transmission path and at the terminal station. Computer memory can store speech with much less distortion than typical analog audio tapes. Time division multiplexing provides a simple and economic way to carry a number of signals simultaneously compared to frequency division multiplexing which requires complex filters for its implementation. Digital switching can also be accomplished faster and cheaper without the problems of analogue cross-talk and mechanical switching. Digitally encoded information such as speech, video computer data, facsimile data can be transmitted over the same communication system. This extended communication system known as ISDN (Integrated services digital network) will provide end-to-end digital connecticity to support a wide range of services.

There are three main factors that characterize a particular speech coder: Speech quality, bit rate and algorithm complexity (hence implementation cost). The three are strongly interrelated. As coder complexity increases, better speech quality can be achieved at lower bit rates. The quality of reproduced speech can be rated as one of four broad categories [9]. (1) Commendary or broadcast quality refers to wide bandwidth (~ 7 KHz) speech with no perceptible distortion; (2) Toll quality equivalent to a "good" analogue sample from the switched telephone network (200-3300 Hz range, signal to noise ratio of more than 30 dB and less than 2-3% harnomic distortion) (3) communication quality which is highly intelligible but noticeably worse compared to Toll quality (4) synthetic quality which is 80-90% intelligible, has substantial degradation (sounds "machine like" and often "buzzy" and suffers from a lack of speaker identifiability. A relationship between these broad categories and the bit rates at which this quality can be achieved are shown in figure (3.2-1). Note that the divisions are only approximate. The Toll quality barrier is likely to move towards 8 Kb/s and the communications quality to pass the 4.8 kb/s barrier and perhaps reach the 2.4 kb/s rate in the near future. The descriptors on the top refer to the particular type of coder used to achieve the specified quality at the range of bit rates shown. In waveform coding systems an attempt is made to preserve the waveform of the speech signal. Usually waveform coders can code equally well a variety of signals. They tend to be robust for a wide range of speaker characteristics and for noisy environments. The waveform coders operating at the lower bit rates exploit various redundancies in the signal and, also, adapt better to the nonstationarity of the source. Source coders (vocoders) on the other hand make no attempt to preserve the signal waveform. During the encoding procedure, at the transmitter, a model for the speech production mechanism is employed and its parameters determined from the speech signal. These parameters are usually split into vocal tract and excitation parameters. A very small set of parameters is usually used to mainly reproduce the short term speech envelope, related to the vocal tract model and provide a crude representation of the fine structure of the short term speech spectrum. Since intelligibility depends primarily on the above features, vocoders find useful applications where a small bandwidth (bit rate) is a strict requirement rather than quality. The quality of vocoder systems is poor, mainly due to the inadequate modelling of the excitation part of the production model. Another distinction that can be made between waveform and source coders is that a waveform coder with uncoded parameters would be indistinguishable from the original whereas a vocoder with unquantized parameters will still give a distorted signal. The third category of coders, that of hybrid coders bridge the quality gap between vocoders and waveform coders (figure 3.2-2). [10]. These coders are similar to vocodes in that a production model is used but, are also similar to waveform coders in that the waveform of the coded speech bears close similarity to the original. For this class of coders the perceptually significant part of the spectrum is "waveform encoded" (i.e. both phase and amplitude are preserved) whereas the rest of the spectrum is "vocoder driven" retaining only its broad features. Hybrid coders have become increasingly important in recent years for toll telephone quality below 16 kb/s and communications quality around 4.8 kb/s for mobile radio applications.

A review of representative coders from each class will be given below with more emphasis on waveform coders and hybrid coders.

3.2.2 Vocoders

3.2.2.1 Channel Vocoders

Channel vocoders obtain a short time amplitude spectrum representation of the speech signal by bandpass filtering, rectification and lowpass filtering [9], an approximation to quadrature envelope detection [11] (figure 3.2-3). The bandpass filters are usually contiguous and the bandwidth may increase with frequency to reflect the ear's frequency resolution in relation to frequency. The aim, at least for the low frequency range, is to isolate each harmonic within one band and hence reproduce the correct amplitude of that harmonic at the synthesizer. When more than one harmonic is present, the channel output will be an average of the amplitudes of the harmonics within that frequency band and therefore the amplitude spectrum at the synthesizer will be somewhat distorted. At higher freqencies, individual harmonic amplitudes cannot be resolved since the ear integrates energy within a critical band and produces an average value, much as the analysis part of the channel vocoder does. Note that the low-pass filters have a high enough cutoff (20 Hz) to convey the necessary loudness fluctuations of each component but also low enough to prevent any roughness forming in any channel, at least from the vocal tract representation. The outputs of the rectifiers (or the low pass filters) are sampled every 20 ms which requires a bandwidth of 1/(2 x 20 ms) = 25Hz. For the excitation, a voiced-unvoiced detector determines whether white noise or a series of pulses is to be used at the synthesizer, and in the latter case a pitch detector determines the spacing between these pulses. Utilizing time and frequency redundancies, the required information can be transmitted at about 1.2 kbs/sec [12]. Variable frame rate transmission [13] can also be used: During quasi-stationary segments of speech spanning more than one frame of side information only the beginning and end frame of the segment need be transmitted and the rest can be derived through interpolation. In addition vector quantization techniques can be used to further compress the required rate.

3.2.2.2 LPC vocoders:

The LPC vocoder is a time domain equivalent to the channel vocoder. Here the vocal tract information (plus the radiation characteristics and the pulse shape of the vocal excitation) is modelled using a p^{th} order linear all pole filter of the form [6]

$$H(z) = \frac{G}{1 - \sum_{i=1}^{p} a_i Z^{-i}} \qquad 3.2\text{-}1$$

The excitation part is similar to that in the channel vocoder, involving a voiced-unvoiced decision and the transmission of pitch (pulse spacing) information for voiced segments. A schematic diagram of the model is given in figure (3.2-4). Since the filter represents the spectral envelope of speech, whereas the excitation represents the fine structure, a multiplication of the two in the frequency domain should produce a signal "close" to the spectrum of the original speech segment the model simulates. In the time domain this process is equivalent to convolution.

Let the excitation be represented by U(Z) and the speech signal by S(Z). The speech signal (or a "close" version of it) can be represented by

$$S(Z) = H(Z)\ U(Z) \qquad 3.2\text{-}2a$$

or

$$S(Z) = \frac{GU(Z)}{1 - \sum_{K=1}^{P} a_K Z^{-K}} \qquad 3.2\text{-}2b$$

and as a convolution in the time domain

$$S(n) = \sum_{K=1}^{P} a_K S\,(n-K) + Gu(n) \qquad 3.2\text{-}2c$$

The coefficients $\{a_K\}$ can be calculated subject to some predetermined criterion. Since all the envelope information must be conveyed in H(Z), one reasonable criterion is to find the $\{a_K\}$ that produce a flat (in an envelope sense) Gu(n) in the frequency domain. It can be easily shown [5] that the coefficients $\{a_K\}$ resulting in a flat spectrum in Gu(n) are the same as the coefficients minimizing the energy in Gu(n) over the speech segment of interest i.e.

$$\text{minimize } \sum_N [Gu(n)]^2 \text{ where N is the range of n} \quad 3.2.3a$$

or

$$\text{minimize } E = \sum_N [S(n) - \sum_{K=1}^{P} a_K S(n-k)]^2 \quad 3.2.3b$$

The values of a_K can be found by setting

$$\frac{\partial E}{\partial a_i} = 0 \text{ for } i=1,\ 2....p \quad 3.2.3c$$

to obtain

$$\sum_N S(n-i)S(n) = \sum_{K=1}^{P} a_k \sum_N S(n-i)S(n-K) \quad 3.2-4$$

Let

$$\phi(i,k) = \sum_N S(n-i)\ S(n-k) \quad 3.2-5$$

then equation 3.2-4 can be written as

$$\sum_{K=1}^{P} a_K\ \phi_n(i,K) = \phi_n(i,0) \qquad i=1,\ 2.....P \quad 3.2-6$$

The solution of the above equation depends on the range N of the sumation for the ϕ terms. A more detailed discussion of the solution will be presented in the section on DPCM coding.

The appropriate pitch period as well as the voiced unvoiced decision can be determined in a variety of ways [1,2,16]. The update of the LPC coefficients is of the order of 20 ms as in the case of channel vocoders. Note that 8 parameters (P=8) provide as good a resolution as

15 channel parameters in the channel vocoder [9]. This is due to the fact that the all-pole model is very suited to the kind of spectrum that vowels possess. Although nasals and frigatives require zeros as well, these are not as important, due to masking effects. For the case of voiced sounds, since the vocal tract filter is kept constant over relatively long periods of time (20 ms) and the excitation is composed from a steady state tone complex (a series of pitch pulses) which results in no amplitude fluctuation in the frequency domain, the quality of the synthesized speech for correctly identified segments is "smooth" sounding and free from roughness. The overall quality of the speech is nevertheless poor. This is because the analyser can sometimes identify a voiced sound to be unvoiced and vice-versa which results in occasional harshness and buzziness in the synthesized speech. Further deteriorations in quality occur due to errors in the estimation of the correct pitch period of the analysed sounds. These effects can degrade substantially the quality of the synthesized speech even when the analyser estimates accurately the excitation parameters for 95% of the time. Various modifications to the excitation model have emerged throughout the years [14, 15]. Other vocoder structures are the homomorphic vocoder [17, 18], the phase vocoder [2,11] the formant vocoder [19] and others [1]. It now seems possible to transmit intelligible speech at rates as low as 200 b/s [20]. Although quantization of parameters in vocoder systems is very important for very low bit rates such as those above we will reserve a discussion on quantization for the following section, on waveform coding.

3.2.3 Waveform Coding

3.2.3.1 PCM

In almost all waveform coders the band limited analogue speech signal X(t) is first sampled at a rate greater or equal the Nyquist rate (i.e. $2f_{max}$ where f_{max} is the higher frequency present in X(t)) to produce a series of samples X(n). At this point the process is reversible since the original analogue band limited signal can be reproduced exactly from its sampled version. The next step in digitizing the signal is to introduce some form of amplitude quantization into the signal. The simplest form of amplitude discretization is pulse code modulation. This is the first method "historically", of converting analogue speech signals into a digital form and is still widely used in commercial digital speech transmission systems. It also serves as a preprocessing stage to more sophisticated algorithms which are geared to operate on digital samples.

After sampling, the amplitude continuous values X(n) are converted to the nearest of a finite set of amplitude levels Y(n). The number of these levels determines the bit rate. The spacing of these reconstruction levels determines the quality of the reconstructed signal for a fixed number of levels. There are generally two types of quantizer characteristics, the mid-riser figure (3.2-5a) and the mid -tread (figure 3.2-5b). Two types of quantization noise arise from this process, one is granular noise, when the amplitude of the sample X(n) falls within the quantizer range and overload or clipping noise when the amplitude X(n) falls beyond the quatizer range, figure (3.2-5). The spacing of the levels determine which of the two types of noise prevails and a choice can be made according to perceptual or objective criteria provided the probability distribution function (Pdf) of the source is known. For a quantizer of L-levels the bit rate is

$$R = \log_2 L \text{ bits/sample} \qquad 3.2\text{-}7$$

Associated with the quantization process is the quantization error q(n) defined as

$$q(n) = y(n) - x(n) \qquad 3.2\text{-}8$$

The most important quantity for comparing the performances of quantizers is the quantitation error variance σ_q^2

$$\sigma_q^2 = E[q(n)]^2 \qquad 3.2\text{-}9a$$

and E denotes expectation

$$= \int_{-\infty}^{+\infty} q^2 P_q(q)\, dq \qquad 3.2\text{-}9b$$

where $P_q(q)$ is the pdf of q(n)

alternatively, since the quantizer characteristic is fixed for a given step size and quantizer type (fig. 3.2-5) q(n) is a function of X(n) and the noise variance is given by

$$\sigma_q^2 = \int_{-\infty}^{+\infty} [y-x]^2\, P_x(x)dx \qquad 3.2\text{-}9c$$

where $P_x(x)$ is the pdf of the input, X(n). It is therefore clear that σ_q^2 depends on the quatizer characteristic (which determines y for any given x) and the Pdf of the input signal.

For a bounded input $|X| \leqslant X_{max}$ and sufficiently small step size Δ, the noise variance, provided there is negligible overload, is given by [21]

$$\sigma_q^2 = \frac{\Delta^2}{12} \qquad 3.2\text{-}10$$

If in addition the range of the quantizer is reduced to just accommodate the signal i.e. the condition of negligible overload is just fulfilled, the noise variance is given by

$$\sigma_q^2 = \frac{1}{3} X^2_{max} 2^{-2R} \qquad 3.2\text{-}11$$

The quantizer performance is usually expressed in the form

$$\sigma_q^2 = \epsilon_*^2 2^{-2R} \sigma_x^2 \qquad 3.2\text{-}12a$$

Note that to convert equation 3.2-11 to 3.2-12 a knowledge about the input signal's pdf is required to relate X_{max} to σ_x.

R can then be expressed as

$$R = \frac{1}{2} \log_2 (\sigma_x^2 / \sigma_q^2) + \frac{1}{2} \log_2 \epsilon_*^2 \qquad 3.2\text{-}12b$$

bits/sample.

The term

$\dfrac{\sigma_x^2}{\sigma_q^2}$ is the signal to quantization noise ratio:

$$\text{SNR (dB)} = 10 \log_{10} \frac{\sigma_x^2}{\sigma_q^2}$$

and ϵ_*^2 is related to the quatizer performance.

For a given value of σ_q^2 the minimum bit rate required, as given by rate distortion theory is of the form

$$\min (R) = \frac{1}{2} \log_2 \left(\frac{\sigma_x^2}{\sigma_q^2} \right) - a_R \text{ bits/sample} \qquad 3.2\text{-}13$$

where a_R (≥ 0) depends on the statistics of the source X

The SNR as derived from equation 3.2-12b is given by

$$\text{SNR (dB)} = 6.02R - C \qquad 3.2\text{-}14$$

C depends upon the particular quantizer and input characteristics.

To minimize σ_q^2 for the unbounded pdf implies a balance between granular and overload noise.

With the constraint of a midrize (symmetric) uniform quantizer of step size Δ, the optimum value of the step size Δ_{opt} in relation to σ_x, the standard deviation of the signal is given in Table (3.2-T1) for various pdfs. The pdfs are shown in figure 3.2-6 and defined in table (3.2-T2). The SNR in dB is also shown in table (3.2-T1).

3.2.3.1a Non Uniform Quantization

For nonuniform pdf uniform quantization is not the optimum solution. Smaller noise variance can be obtained using nonuniform quantization (i.e. a step size that changes with X). Smaller step sizes can be used where the probability of occurrence of X is high at the expense of larger step sizes where the pdf of X is relatively low. The procedure of using a nonuniform quantizer is equivalent to compressing the signal X using a nonuniform compressor characteristic C(X), quantizing the compressed signal C(X) using a uniform quantizer, and expanding the quantized signal using the inverse characteristic $C^{-1}(X)$ to that of the compressor. The compressor characteristic is refered to as the companding law. The procedure is shown in figure 3.2-7.

Pdf-optimized nonuniform quantizers can also be designed [22, 23]. The minimum mean-squared-error (noise) min $\{\sigma_q^2\}$ quantizer is usually called the Max or Max-Lloyd quantizer. Note that the optimization with regard to pdf shape includes a match of quantizer to input variance σ_x^2. Also the input and quantization error are correlated and the variance of the output of the quantizer is always less than that of the input variance:

$$\sigma_y^2 = \sigma_x^2 - \min \{\sigma_q^2\} \qquad 3.2\text{-}15$$

The optimum decision values X_j and reconstruction values Y_j for various pdfs for pdf-optimized nonuniform quantizers are given in table 3.2-T3 and the corresponding quantizer performance in table 3.2-T4. The effects of quantizer mismatch either in variance or pdf shape can be found in [21].

3.2.3.1b Logarithmic quantization

Pdf-optimized quantizers are also matched to a particular input variance (a variance of 1 in Table 3.2-T3). In situations such as speech coding the exact value of the input variance is not known in advance; and in addition it tends to change with time. In such situations a signal to quantization noise ratio that is constant over a broad range of input variances can be obtained by using a logarithmic companding law. This is illustrated in figure (3.2-8). The vertical denotes the SNR whereas the horizontal gives the input variance of a laplacian input relative to its boundary value of X_{max}. The dashed line gives the performance of the optimum laplacian quantizer for a particular σ_x/X_{max} ratio. It can be seen that the optimum quantizer has a higher maximum SNR but a much smaller dynamic range than the logarithmic quantizer.

There are two wide-spread compression characteristics for logarithmic quantizers:

The first is the A-law-companding given by

$$C(X)=\begin{cases} \dfrac{A|x|}{1+\log_e A}\,\mathrm{sgn}(X) \quad , & 0\leqslant \dfrac{|X|}{X_{max}} \leqslant \dfrac{1}{A} \\[2ex] X_{max}\,\dfrac{1+\log_e(A|X|/X_{max})}{1+\log_e A}\,\mathrm{sgn}(X) \; , & \dfrac{1}{A} < \dfrac{|x|}{X_{max}} \leqslant 1 \end{cases} \qquad 3.2\text{-}16$$

For the European PCM standard A = 87.56 and $SNR_{a-law}(dB)$ = 6.02R-9.99 3.2-17

The other compounding law is the μ-law defined as

$$C(X) = X_{max}\,\frac{\log_e(1+\mu\,\dfrac{|x|}{X_{max}})}{\log_e(1+\mu)}\,\mathrm{sgn}\,X \qquad 3.2\text{-}18$$

For the North-American PCM standard μ = 255

$$SNR_{\mu-law} = 6.02R - 10.1 \qquad 3.2\text{-}19$$

Commercial PCM systems are based on the use of 8 bits/sample. With a speech sampling rate of 8KHz the transmission rate is 64 kb/s. An SNR of nearly 34 dB is maintained over a range of input signal variation of around 30 dB. In practice a piecewise linear approximation is made of the companding law to convert to and from 8 bit log PCM to 12 or 13 bit linear PCM which are considered as equivalent. PCM is the best established, the most implemented and the most applied of all digital coding systems. This is due to the fact that it is the earliest developed system, it is simple, instantaneous (or near instantaneous) and it is not signal specific. Finally, PCM coding serves as a front-end preprocessing stage to most other more complex waveform coders.

3.2.3.1c Entropy-Coding

In the above discussion, non-uniform quantization was used to take advantage of the nonuniform pdf of the source (input signal). Each quantized sample is then assigned a fixed number of bits as in log-PCM.

Another way of taking advantage of a nonuniform Pdf is entropy coding [21]. In Entropy coding a uniform quantizer can be used with sufficient range to avoid overload and a step-size dependent on the required bit rate. The number of levels of the uniform quantizer are usually much larger than the quantizer of equivalent performance that does not employ Entropy coding. In order to keep the bit rate low, different number of bits is assigned to each level: Highly probable levels i.e. levels within the region where the pdf of X is high are assigned short codewords (small number of bits) whereas the outsider less frequent values of X are assigned longer codewords, with the length increasing with decreasing probability of occurrence of that level. Note that the smallest bit rate that can be achieved here is again 1 bit/sample as in ordinary PCM since this is the smallest codeword.

If Entropy coding is now used on sequences of samples (rather than on single outputs) a bit rate lower than 1 bit/sample can be achieved. Linear redundancies are usually removed through linear filtering prior to entropy quantization. This is in order to avoid processing long blocks of data and calculating joint probabilities to construct the appropriate quantizer codebook. An interesting feature of entropy coding is that a midtread characteristic can result in a significant reduction in distortion compared to a midrize quantizer, if the source has values close to zero with high probabilities. Finally a buffer must usually be employed since both the source and the channel operate at a constant rate whereas entropy coding necessitates variable rate coding.

3.2.3.1d Representation of noise

For additive input independent noise q the following equation relating input, output and noise variances,

$$\sigma_y^2 = \sigma_x^2 + \sigma_q^2 - 2E\{xq] \qquad 3.2\text{-}20$$

can be simplified to :

$$\sigma_y^2 = \sigma_x^2 + \Delta^2 /12 \text{ since } E[xq]=0 \qquad 3.2\text{-}21$$

(figure 3.2-9a)

For the case of a pdf optimized quantizer though

$$E[xq] = \min\ \{\sigma_q^2\} \text{ and } \sigma_y^2 = \sigma_x^2 - \min\ \{\sigma_q^2\} \qquad 3.2\text{-}22$$

[21]

A model for this kind of noise is one with a less than unity gain component a_g and an additive uncorrelated component n. (figure 3.2-9b). Forcing n to be additive and input independent i.e.

$$E(xn) = 0, \; E(xq) = E[X(X-a_g X-N)] = \sigma_x^2 (1-a_g)$$

and from 3.2-22

$$a_g = 1 - \min \{\sigma_q^2\}/\sigma_x^2 \qquad 3.2\text{-}23$$

For low bit rates therefore, the errors in quantization can be considered to belong to two different classes. One caused by a change in level, which has little perceptual significance unless the change exceeds 3 dB, and one due to uncorrelated noise which is the main source of perceptual degradation. This is one of the reasons why the SNR is not a very good measure for low bit rates.

3.2.3.1e Adaptive quantization

We have seen that a log-companding quantizer is quite successful in reducing the required bit rate for Toll quality speech from 12 bits to 8 bits per sample. The nonstationarity of the speech signal was taken into account by providing smaller step sizes for low levels of speech and increasing the step size for higher levels. Such a quantizer is time-invariant. A better performance can be achieved using a time varying or adaptive quantizer. This type of quantizer is essential when the required bit rate falls below about 5 bits/sample.

In an adaptive quantizer the step size Δ changes with time so that a value close to optimum is available for each sample. From previous sections the optimum step size is proportional to σ_x, the proportionality factor depending on the input pdf and the bit rate. Therefore

$$\Delta(n) = a\sigma_x(n) \qquad 3.2\text{-}24$$

where a is a constant.

For a nonstationary input σ_x is variable and the determination of an appropriate step size involves the continuous estimation of σ_x. Two ways have been devised for tracking the input level in terms of σ_x. One system is denoted by AQF (adaptive quantization with forward estimation) and the other by AQB (adaptive quantization with backward estimation). Forward estimation is based on unquantized samples X(n). It is therefore unaffected by quantization noise. It also creates an additional information (side information) that has to be transmitted to the receiver. This information requires relatively small bit rates compared with the signal rate and therefore permits special protection of step size information from channel noise resulting in a more robust coder. Backward estimation is based on past quantizer output samples y(n) and therefore is not as reliable an estimate as that obtained through AQF since it is affected by quantization noise. It does not require any additional side information since the past samples y(n) are also available at the receiver. This also makes AQB more susceptible to channel errors. A diagram of the two methods is shown in figure 3.2-10. AQF is necessarily block adaptive since bit rate restrictions do not allow an update of σ_x on a sample by sample basis. This also results in a necessary block delay (usually around 16 ms) for the calculation of σ_x. For this reason AQF is also refered to as block companding. The adaptive quantizers need not be uniform but can have nonuniform characteristics optimized for the pdf in the neighbourhood of the current sample (usually gaussian for speech inputs).

For the AQF case, step size adaptations can follow the input variance via:

$$\Delta(n) = a\sigma_x(n);\ \sigma_x(n) = [\frac{1}{N}\sum_{i=0}^{N} X^2(n+i)]^{1/2} \qquad 3.2\text{-}25$$

The estimate is usually transmitted once every N samples.

For AQB the estimate also involves a number of samples i.e. a time window is applied, through which the variance is estimated. This estimate can now be updated on a sample by sample basis since no transmission of the step size information is involved here. By using an exponential window the calculation burden can be reduced and the estimate can be given by

$$\sigma_y^2(n) = a\sigma_y^2(n-1)+(1-a)y^2(n-1) \qquad 3.2\text{-}26$$

for a ≈ 0.9 we have instantaneous adaptation whereas for a ≈ 0.99 we have syllabic adaptation, figure (3.2-11).

The above formula implies that effective adaptation can be realized with an explicit memory of only one sample y(n-1) together with the quantizer history condensed into a slowly varying parameter $\sigma_y^2(n-1)$ or equivalently of the corresponding step size $\Delta(n-1)$. This leads to adaptive quantization with one word memory [25]. For a midrize quantizer, if the latest output level is

$$y(n-1) = H(n-1)\Delta(n-1)/2;\ \pm H(n-1)=1,3,5....2^{R-1}$$
$$R \geqslant 2$$
$$3.2.27a$$

the adaptation logic derives the next step size $\Delta(n)$ as the produce of $\Delta(n-1)$ and the multiplier M(.) which is a time invariant function of the latest magnitude index $|H(n-1)|$: i.e.

$$\Delta(n) = M(|H(n-1)|)\Delta(n-1) \qquad 3.2\text{-}27b$$

Various constraints and modifications to the above equation are necessary for a practical system [25,26,27,28].

3.2.3.2 Differential PCM (DPCM)

Differential coding or predictive coding systems utilize waveform redundancy in the time-domain with the corresponding reduction in bit rate for a specified quality of digitization. In general the quantizer input in a DPCM coder is a prediction error or difference signal

$$d(n) = X(n) - \hat{X}(n) \qquad 3.2\text{-}28$$

where $\hat{X}(n)$ is a prediction of $X(n)$. In order that the transmitter and receiver parts of the DPCM system track and reconstruct input waveforms in synchrony, it is essential that the prediction $\hat{X}(n)$ depends on previous quantized values $Y(n) = \tilde{X}(n)$ rather than unquantized inputs $X(n)$. Incorporation of this property leads to the closed loop or feedback around quantizer structure of figure (3.2-12).

The DPCM coder uses linear prediction in the form

$$\hat{X}(n) = \sum_{j=1}^{P} a_j \tilde{X}(n-j) \qquad 3.2\text{-}29$$

where

$A = \{a_j\}$, $j=1, 2,.....P$ is a set of predictor coefficients

The basic equations describing DPCM, from figure 3.2-12:

$$d(n) = X(n) - \hat{X}(n) \qquad 3.2\text{-}30$$

$$u(n) = d(n) - q(n) \qquad 3.2\text{-}31$$

$$y(n) = \hat{X}(n) + V(n) \qquad 3.2\text{-}32$$

y(n) is the coder output, d(n) the prediction error, u(n) the quantized prediction error and v(n) the receiver version of u(n).

Combining 3.2-29 with 3.2-30

$$X(n) = \sum_{j=1}^{p} a_j \tilde{X}(n-j) + d(n) \qquad 3.2\text{-}33$$

for fine quantization $X(n-j) \approx \tilde{X}(n-j)$ equation 3.2-33 is very similar to equation 3.2-2c in the section on LPC vocoders with d(n) replaced by Gu(n) of that section. Due to the similarity, the predictor described by 3.2-29 is refered to as an all pole predictor. Note that the coder of figure 3.2-12 can be regarded as a generalized quantizer whose center-point keeps getting shifted to the latest value of $\hat{X}(n)$. This shifting aligns the quantizer with the amplitude range most likely to be occupied by X(n) and enables the encoder to use finer quantization than in a PCM situation for a given number of quantization levels. Quantization shifting implies predictability of X(n). Clearly in the case of an uncorrelated input X(n), the best estimate of X(n) is the unconditional mean value (usually zero) and therefore no gain due to a DPCM structure can be obtained.

With error free transmission of u(n) (i.e. u(n) = V(n)) the prediction input at the transmitter is

$$\tilde{X}(n) = \hat{X}(n) + u(n) = \hat{X}(n) + V(n) = y(n) \qquad 3.2\text{-}34$$

The reconstruction error r(n) for sample n is given by

$$r(n) = X(n) - y(n) \qquad 3.2\text{-}35$$

From equations 3.2-35 and 3.2-30, 3.2-32, assuming error free transmission

$$r(n) = \hat{X}(n) + d(n) - [\hat{X}(n) + u(n)]$$

$$= d(n) - q(n) \qquad 3.2.36a$$

and from equation 3.2-31

$$r(n) = q(n) \qquad 3.2.36b$$

Equation 3.2-36 reflects an important property of the "closed loop" predictive coding scheme of figure 3.2-12: The quantization noise does not accumulate (i.e. r(n) does not depend on previous quantization noise samples q(n-j)). Note that in the derivation of equation 3.2-36 equation 3.2-29 was not used and therefore 3.2-36 holds for any predictor structure H.

It follows that the mse (mean-square-error) performance of DPCM is described by the following equations

$$\sigma_r^2 = \sigma_q^2 \qquad 3.2\text{-}37$$

$$\sigma_q^2 = \epsilon_*^2 \, 2^{-2R} \, \sigma_d^2 \qquad 3.2\text{-}38$$

(by analogy to equation 3.2-12a)

Let $\sigma_x^2/\sigma_d^2 = G_p$ 3.2-39

From the above three equations the reconstruction error is given by

$$\sigma_r^2(\text{DPCM}) = (\epsilon_*^2 \, 2^{-2R} \, \sigma_x^2) \, \frac{1}{G_p} \qquad 3.2\text{-}40$$

The reconstruction error in PCM is given by

$$\sigma_r^2 \, (\text{PCM}) = \epsilon_*^2 2^{-2R} \, \sigma_x^2 \qquad 3.2\text{-}41$$

Note that ϵ_*^2, the quantizer performance need not be the same in the last two equations. In the case of speech signals, the statistics of X(n) and d(n) (i.e. the pdf) are similar in the case of linear prediction with DPCM and logarithmic quantization with PCM [21] and the two quantizer performances have a ratio near to unity. In this case

$$G_p \frac{\sigma_x^2}{\sigma_r^2 \text{ (PCM)}} = \frac{\sigma_x^2}{\sigma_r^2 \text{ (DPCM)}} \qquad 3.2.42a$$

or

$$\text{SNR (DPCM)}_{dB} = \text{SNR (PCM)}_{dB} + 10\log_{10} G_p \qquad 3.2.42b$$

G_p is typically greater than one. In view of the 6 dB-per bit SNR results of equation 3.2-14, DPCM coding provides a $(10\log_{10}G_p)/6$ bit advantage over PCM. The factor G_p is usually referred to as prediction gain for obvious reasons. Note that the above simplified formula 3.2-42, does not fully describe the performance of DPCM, since the effects of quantizer performance and quantization error feedback (i.e. prediction from noisy samples) have an effect on the resulting SNR. Predictor design usually aims to maximise G_p.

The maximum value of predictor gain $^{N}G_p$ is achieved as $N \rightarrow \infty$. This value provides an upper bound on prediction gain for linear predictors and is equal to the reciprocal of a spectral flatness measure:

$$\max_{N \rightarrow \infty} \{^{N}G_p\} = (\gamma_x^2)^{-1} \qquad 3.2-43$$

where γ_x^2 is defined as

$$\gamma_x^{\,2} = \frac{\exp[\frac{1}{2\pi} \int_{-\pi}^{+\pi} \log_e S_{xx}(e^{jw})\, dw]}{\frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(e^{jw})\, dw} \qquad 3.2\text{-}44$$

and $S_{xx}(e^{jw})$ is the power spectral density of the input variable X(n). Note that

$$\sigma_x^{\,2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} S_{xx}(e^{jw})dw \qquad 3.2\text{-}45$$

is the arithmetic mean of the (power) spectrum and

$$\eta_x^{\,2} = \exp[\frac{1}{2\pi} \int_{-\pi}^{+\pi} \log_e S_{xx}(e^{jw})dw] \qquad 3.2\text{-}46$$

is the geometric mean of the spectrum.

By definition

$$0 \leqslant \gamma_x^{\,2} < 1 \qquad 3.2\text{-}47$$

with $\gamma_x^{\,2} = 1$ for a white noise process

3.2.3.2a The linear predictor of order P

As mentioned earlier the linear predictor is defined by

$$\hat{X}(n) = \sum_{j=1}^{P} a_j y(n-j) = \sum_{j=1}^{P} a_j [X(n-j)-q(n-j)] \qquad 3.2\text{-}48$$

The weighting factors a_j are the predictor coefficients.

The term:

$$QEF(n) = \sum_{j=1}^{P} a_j q\ (n-j) \qquad 3.2\text{-}49$$

represents the quantization error feedback: This term is ignored when designing the predictor (i.e. finding the coefficients $\{a_j\}$), simply because the values q(n) are not known at the design stage. The effect of the terms of 3.2-49 is small for fine quantization.

From equation 3.2-37 and 3.2-38, the reconstruction error variance is proportional to the prediction error variance. the criterion for designing a predictor is therefore to minimize the prediction error variance σ_d^2 from equations 3.2-30, 3.2-48 and the assumption of negligible QEF(n)

$$d(n) = X(n) - \sum_{j=1}^{P} a_j X\ (n-j) \qquad 3.2\text{-}50$$

and we aim to minimize

$$\sigma_d^{\,2} = \sum_{N} \{X(n) - \sum_{j=1}^{P} a_j x(n-j)\}^2 \qquad 3.2\text{-}51$$

Note that this formula is identical to formula 3.2-3 and therefore its solution is of the form

$$\sum_{k=1}^{P} a_k \phi_n\ (i,k) = \phi_n(i,0) \qquad i=1,2.\dots.P \qquad 3.2\text{-}52$$

for

$$\phi(i,k) = \sum_{N} X(n-i)X(n-k) \qquad 3.2\text{-}53$$

In the case of a fixed (nonadaptive predictor) the terms $\phi(i,k)$ can be replaced by the long term autocorrelation coefficients $R(|i-k|)$ and equation 3.2-52 can be written as

$$\phi(i,k) = \sum_{k=1}^{p} a_K R(|i-k|) = R(i) \quad 1 \leqslant i \leqslant p \qquad 3.2\text{-}54$$

The prediction gain G_P is a function of the predictor order P but, for the case of a fixed predictor its value saturates for predictor orders greater than 2 or 3, figure (3.2-13) [21].

Low complexity DPCM systems employ fixed predictors and adaptive quantization [26]. Figure (3.2-14) compares reconstruction error spectra in log-PCM and DPCM-AQB speech coders of identical SNR. Subjectively the DPCM noise spectrum is the prefered shape. This is undoubtedly due to the effect of masking of the noise by the speech signal. Higher masking is produced where the speech signal power is higher i.e. over the low frequency region. The quality of low bit rate DPCM can be enhanced by adaptive postfiltering procedures which capitalize on the fact that the short-time speech cutoff frequency is often less than the nominal 3.4 KHz [29].

3.2.3.2b Adaptive prediction

Input statistics such as autocorrelation functions and probability density functions are time varying in the case of nonstationary signals and as a consequence, best predictor designs for inputs such as speech should be time varying, or adaptive as well.

As in the case of sample quantization, adaptation can be achieved in a forward or backward mode. Figure 3.2-15 shows the two different structures. For forward adaptation M input samples are buffered and a set of prediction coefficients is calculated which is optimum for the buffered speech segment. For speech sampled at 8 KHz a good choice for the buffer length is 16 ms whereas a prediction order of around 10 is adequate. The depedence of gain G_P upon prediction order is shown in figure (3.2-16). Note that prediction gains saturate much slower with increasing predictor order than in the nonadaptive case (figure 3.2-13). Further comparisons between adaptive and nonadaptive prediction can be seen in figure (3.2-17): With adaptive prediction G_P is always positive whereas in the nonadaptive case G_P is sometimes negative.

A comparative study of digital waveform coding schemes involving PCM, DPCM and ADPCM was carried out by Noll [30]. He considered the following systems:

1. μ = 100 log PCM with $X_{max} = 8\sigma x$ (PCM)
2. Adaptive PCM (optimum Gaussian quantizer) with feed-forward control. (PCM-AQF)
3. Differential PCM with first order fixed prediction and adaptive Gaussian quantizer with feedback control. DPCM1-AQB.
4. Adaptive DPCM with first order adaptive predictor and adaptive Gaussian quantizer with feed-forward control of both the quantizer and the predictor (window length 32) ADPCM1-AQF.
5. Adaptive DPCM with fourth order adaptive predictor and adaptive Laplacian quantizer, both with feed-forward control (window length 128) ADPCM4-AQF
6. Adaptive DPCM with twelfth order adaptive predictor and adaptive Gamma quantizer, both with feed-forward control (window length 256) (ADPCM12-AQF).

In the above systems the sampling rate was 8 KHz and the quantizer word length ranged from 2 bits/sample to 5 bits/sample. Thus the bit rate range is from 16 kb/s to 40 kb/s. The signal to quantization noise ratios are shown in figure 3.2-18. The lowest curve corresponds to the use of a 2-bit quantizer and moving upward from one curve to the next corresponds to adding one bit to the quantizer word length. Note that the curves are displaced from one another by roughly 6 dB. The perceived quality of ADPCM coded speech is better in comparison to PCM by a greater extent than the SNR values would suggest.

A comparison is given in table 3.2-T5. ADPCM here is scheme 3 whilst PCM is scheme one [26].

3.2.3.2C Adaptive predictor design

For the adaptive case equations 3.2-52 and 3.2-53 still hold, but N, the range of summation is no longer large but of the order of 20 ms. There are basically two approaches to this short-time analysis procedure, the autocorrelation method and the covariance method. The assumption implied in equation 3.2-50 still holds. i.e. quantization noise is neglected.

3.2.3.2C1 The Autocorrelation method [5, 31]

In this method, the limits of summation assume that the signal is zero outside the interval $0 \leq n \leq N-1$. In this case the prediction error d(n) will be nonzero over the interval $0 \leq n \leq N-1+P$. Therefore the quantity to be minimized is

$$\sigma_d^2 = \sum_{n=0}^{N+P-1} d^2(n) \qquad 3.2\text{-}55$$

From equation 3.2-50, the prediction error will be large at the beginning of the interval because one is trying to predict the signal from zero values (the prediction error over this region will be similar to the signal itself). Likewise the error will be large at the end of the interval since one is trying to predict zero values from nonzero values. For this reason a tapering window is usually applied to the data instead of the rectangular window implied above. Equations 3.2-52 and 3.2-53 can be written as

$$\phi\ (i,k) = \sum_{n=0}^{N+P-1} X(n-i)X(n-k) \qquad \begin{matrix} 1\leqslant i\leqslant P \\ 0\leqslant K\leqslant P \end{matrix} \qquad 3.2\text{-}56a$$

or

$$\phi(i,k) = \sum_{n=0}^{N-1-(i-k)} X(n)\ X(n+i-k) \qquad \begin{matrix} 1\leqslant i\leqslant P \\ 0\leqslant k\leqslant P \end{matrix} \qquad 3.2\text{-}56b$$

equation 3.2-56b can be taken to represent a short time autocorrelation function:

$$\phi\ (i,k) = R\ (i-k) \qquad 3.2\text{-}57$$

where

$$R(k) = \sum_{n=0}^{N-1-k} X(K)X(m+K) \qquad 3.2\text{-}58$$

R(k) is an even function hence

$$\phi(i,k) = R(|i-k|) \qquad \begin{matrix} i=1,2\ldots P \\ k=0,1\ldots.P \end{matrix} \qquad 3.2\text{-}59$$

Equation 3.2-52 can then be expressed as

$$\sum_{K=1}^{P} a_k R(|i-k|) = R(i) \qquad 1\leqslant i\leqslant P \qquad 3.2\text{-}60$$

which is identical in form to 3.2-54 although R(k) is now a short-term autocorrelation function defined by 3.2-58, and X(K) is a windowed signal. The resulting short term prediction error variance can be shown to be

$$\sigma_d^{\,2} = R(0)-\sum_{k=1}^{P} a_k R(K) \qquad 3.2\text{-}61$$

which can be used as a scaling factor in an AQF structure.

The matrix equation corresponding to equation 3.2-60 has a Toeplitz structure which results in a fast computation of the predictor coefficients. This can be performed with the Levinson and Robinson algorithms [5] although the most efficient method for solving equations 3.2-60 is Durbin's recursive procedure [31] which can be stated as follows:

$$E^{(o)} = R(o) \qquad 3.2\text{-}62$$

$$K_i = \{R(i) - \sum_{j=1}^{i-1} a_j^{(i-1)} R(i-j)\}/E^{(i-1)} \qquad 1 \leqslant i \leqslant P \qquad 3.2\text{-}63$$

$$a_i^{(i)} = K_i \qquad 3.2\text{-}64$$

$$a_j^{(i)} = a_j^{(i-1)} - K_i a_{i-j}^{(i-1)} \qquad 1 \leqslant j \leqslant i-1 \qquad 3.2\text{-}65$$

$$E^{(i)} = (1-k_i^2)E^{(i-1)} \qquad 3.2\text{-}66$$

The above equations are solved recursively for i = 1,2....P and the solution is given as

$$a_j = a_j^{P} \qquad 1 \leqslant j \leqslant P \qquad 3.2\text{-}67$$

Note that if the autocorrelation coefficients R(i) are replaced by a set of normalized autocorrelation coefficients r(k) i.e. r(k)=R(k)/R(0) then the solution to the matrix equation remains unchanged. Another important fact about the autocorrelation solution is that the resulting filter structure is guaranteed to be stable.

Fast calculation of the autocorrelation coefficients is also possible [5,32].

3.2.3.2.C2 The covariance method [6]

For the covariance method the prediction error variance is minimized over the entire length of the speech segment under consideration:

$$\text{minimize } \sigma_d^2 = \sum_{n=0}^{N-1} d^2(n) \qquad 3.2\text{-}68$$

then ϕ (i,k) becomes

$$\phi(i.k) = \sum_{n=0}^{N-1} X(n-i)X(n-k) \qquad \begin{matrix} 1\leqslant i\leqslant P \\ 0\leqslant k\leqslant P \end{matrix} \qquad 3.2\text{-}69a$$

and with a change of summation index

$$\phi(i,k) = \sum_{m=-i}^{N-1-i} X(i)X(n+i-k) \qquad \begin{matrix} 1\leqslant i\leqslant P \\ 0\leqslant k\leqslant P \end{matrix} \qquad 3.2\text{-}69b$$

or

$$\phi(i,k) = \sum_{m=-k}^{N-k-1} X(n)X(n+k-i) \qquad \begin{matrix} 1\leqslant i\leqslant P \\ 0\leqslant k\leqslant P \end{matrix}$$

Note that for the calculation of $\phi(i,k)$, values of X(n) outside the interval $0\leqslant n\leqslant N-1$ are required and used. In this case no need for tapered windowing arizes since the necessary values outside the interval are made available.

The resulting equation for solution is

$$\sum_{k=1}^{P} a_k\ \phi(i,k) = \phi(i,0) \qquad i=1,2,....P \qquad 3.2\text{-}70$$

The matrix in the corresponding matrix equation describing equation 3.2-70 is a symmetric positive definite matrix. The resulting method of solution is called the Cholesky decomposition [2,5] which provides a fast means of obtaining the predictor coefficients from (3.2-70).

Fast procedures for calculating the covariance terms $\phi(i,k)$ can be found in [5,33]. Although the covariance solution does not guarantee stability, for large N, the solutions are always stable, since, in this case, the solution is virtually identical to the autocorrelation solution. An improvement of the covariance method is the stabilized covariance formulation [35, 36]. This modification acknowledges the fact that, the input to the quantizer consists of both the prediction error as given by equation 3.2-50, but also the quantization error feedback term QEF of equation 3.2-49. Therefore, the total power E_q at the input to the quantizer is the sum of the powers in the prediction error d(n) (E_p) and the filtered noise QEF(n) (E_f). Assuming uncorrelated prediction error and quantization noise:

$$E_q = E_p + E_f \qquad 3.2\text{-}71$$

The power of QEF(n) is determined both by the power in the quantizer error q(n) (eq. 3.2-31 and 3.2-48) and the power gain of the predictor filter P(Z) defined by

$$P(Z) = \sum_{k=1}^{P} a_k Z^{-k} \qquad 3.2\text{-}72$$

Assuming a white quantization noise the power gain of the filter in 3.2-72 is given by

$$G = \sum_{k=1}^{P} a_k^2 \qquad 3.2\text{-}73$$

The power gain can often exceed 200. Such a high power gain causes excessive feedback of the noise power to the quantizer input, for coarse quantizers, resulting in poor performance of the coder. This effect also causes instabilities in the coder system, in spite of the fact that the filter itself may be stable [37]. The reason for the high power gain is is follows: The spectrum of the filter 1-P(Z) is the reciprocal of the envelope of the speech spectrum. The lowpass filter used in the A/D conversion of the speech signal forces the reciprocal spectrum and thus |1-P(Z)| to assume a high value in the vicinity of the cutoff frequency of the filter. The power gain which is equal to the integral of the power spectrum |1-P(f)| with respect to the frequency variable f is also large. This can be corrected by adding to the covariance matrix terms in equation 3.2-69 a correction term corresponding to equivalent covariance terms from a high passed white noise, where the high-pass filter approximates the filter, complimentary to the low pass filter used in the sampling process [35, 36].

The resulting filter H(Z) (as defined by equation 3.2-1) is shown in figure 3.2-19 together with the standard covariance solution. The equivalent reduction in power gain can be seen in figure 3.2-20, and the corresponding reduction in amplitudes of the predictor coefficients in figure 3.2-21. This last figure suggests that the high frequency correction will facilitate the quantization of the prediction coefficients. Another modification to the covariance method is to correct the effect that the position of the analysis frame has upon the prediction error and prediction coefficients. This results in windowing the product defined by $\phi(i,k)$ in 3.2-69 prior to the solution of the matrix [38].

3.2.3.2.C3 Lattice solutions - the Burg method

The lattice solutions evolve from a recursive algorithm for the calculation of the predictor filter without the need for explicit calculation of the correlation (or covariance) matrix.

As in the Durbin algorithm the coefficients $\{a_j^i,$ $j=1, 2, \ldots i\}$ are the coefficients of the ith order optimum linear predictor. Using these coefficients the ith prediction error filter (or inverse filter) can be defined:

$$A^i(Z) = 1 - \sum_{k=1}^{i} a_k^i Z^{-k} \qquad 3.2\text{-}73$$

This filter has the prediction error as its output when the input to the filter is the speech signal itself:

$$d(n) = X(n) - \sum_{k=1}^{i} a_k^i X(n-k) \qquad 3.2\text{-}74$$

It is also possible to define another error signal, the backward prediction error which is defined by

$$b^i(n) = X(n-i) - \sum_{i=1}^{i} a_k^i X(n+K-i) \qquad 3.2\text{-}75$$

which is the prediction error resulting from attempting to predict X(n-i) from the i samples of the input {X(n-i+k), k=1, 2,...i} that follow X(n-i). The i samples involved in the prediction implied by equation 3.2-75 are the same ones involved in the prediction of X(n) in equation 3.2-74. It can be shown that [2]

$$d^i(n) = d^{i-1}(n) - k_i b^{(i-1)}(n-1) \qquad 3.2\text{-}76$$

and

$$b^{i}(n) = b^{i-1}(n-1) - k_{i}d^{i-1}(n) \qquad 3.2\text{-}77$$

The above two equations define the forward and backward prediction errors for an ith order predictor in terms of the corresponding prediction error of an (i-1)th order predictor. A zeroth order predictor implies

$$d^{\circ}(n) = b^{\circ}(n) = X(n) \qquad 3.2\text{-}78$$

equations 3.2-76 and 3.2-77 can be depicted by the flow graph of figure 3.2-22. Such a structure is called a lattice network.

The k_i parameters can be computed directly from the backward and forward errors [2,39]:

$$K_{i} = \frac{\sum_{n=0}^{N} d^{(i-1)}(n)b^{(i-1)}(n-1)}{\{\sum_{n=0}^{N-1}(d^{i-1}(n))^{2}\sum_{n=0}^{N-1}(b^{i-1}(n-1))^{2}\}^{½}} \qquad 3.2\text{-}79$$

The expression describes the degree of correlation between the forward and backward prediction error. For this reason the K_i parameters are usually referred to as the partial correlation coefficients or PARCOR coefficients.

If equation 3.2-79 replaces equation 3.2-63 in the Durbin algorithm identical results obtain for the predictor coefficients.

A new approach that relates to the above has been developed by Burg [2, 40]. This algorithm minimizes the sum of the mean squared backward and forward prediction errors:

$$\text{minimize} \quad \sum_{n=0}^{N-1} [(d^{i}(n))^{2} + (b^{i}(n))^{2}] \qquad 3.2\text{-}80$$

The solution for the k_i coefficients is

$$K_i = \frac{\sum_{n=0}^{N-1} [d^{i-1}(n)\ b^{i-1}(n-1)]}{\sum_{n=0}^{N-1} [d^{i-1}(n)]^2 + \sum_{n=0}^{N-1} [b^{i-1}(n-1)]^2} \qquad 3.2\text{-}81$$

It can be shown [5] that the above solution always yields a stable filter i.e.

$$-1\leq k_i\leq 1 \qquad 3.2\text{-}82$$

equation 3.2-81 can be used to replace equation 3.2-63 in the Durbin algorithm and hence obtain a different set of predictor coefficients.

The Burg algorithm performs well even with small block sizes N and provides better results than the autocorrelation method [41]. It yields a stable filter even without the use of a window.

A complexity comparison between the three methods of obtaining the predictor filter, the Cholesky Decomposition, the Durbin Method and the Burg Method is given in Table 3.2-T6.

3.2.3.2d Split Band LPC

One way to reduce computation is to split the frequency region into a number of bands. Since the complexity is usually proportional to the product of the number of samples N in the frame and the predictor order, which are both reduced in a split band situation, split band systems can model the envelope with less complexity [46]. The above procedure can also be used to give different accuracy to the model for different frequency regions through selective linear prediction [42]. This procedure is beneficial for wide bandwidth speech (~8KHz). Different accuracy for different frequency regions can also be achieved without explicit band splitting of the input signal, [43-45, 47, 48]. Although variations do exist, the general procedure is as follows: [45]

a) window the signal and compute its spectrum
b) warp the spectrum as desired
c) Take the Fourier Transform of the warped spectrum to get the autocorrelation R(i)
d) Solve for the predictor parameters from the normal equations

$$\sum_{K=1}^{P} a_k R(|i-k|) = R(i) \qquad 1 \leqslant i \leqslant P \qquad 3.2\text{-}82$$

These predictor parameters correspond to a warped spectrum and, therefore cannot be used for processing. Dewarping must be performed to obtain a new set of prediction coefficients as follows:

a) Using the predictor coefficients calculated from 3.2-82 calculate the all-pole spectrum from

$$P(w) = \frac{1}{|1 + \sum_{k=1}^{P} a(k) e^{-jkw}|^2} \qquad 3.2\text{-}83$$

b) dewarp this spectrum using the inverse of the function used in the original warping.
c) Take the Fourier Transform of the dewarped spectrum to obtain the corresponding autocorrelation function
d) Equation 3.2-82 can now be used to obtain the new prediction coefficients.

If the spectrum is warped such that high frequencies are compressed relative to low frequencies, low frequencies are better matched than high frequencies since the latter are compressed. This is because, in the warped frequency domain, spectral matching is uniform. Perceptual functions can be used for warping, such as the critical band rate function. This was not found to be satisfactory in [45] perhaps due to the fact that the perceptually good match to the speech signal is a bad model for the production process of speech. A "compromize" warping function can be used for wideband speech.

The above procedure is also useful for unscrambling Helium Speech [43], and in speech analysis [47, 48]. The prediction equations 3.2-50 can be written in the Z-transform domain as

$$D(Z) = X(Z) - \sum_{k=1}^{P} a_k X(Z) Z^{-k} \qquad 3.2\text{-}84a$$

or

$$D(Z) = X(Z) \left[1 - \sum_{k=1}^{P} a_k Z^{-k}\right] \qquad 3.2\text{-}84b$$

and

$$X(Z) = \frac{D(Z)}{1 - \sum_{k=1}^{P} a_k Z^{-k}} \qquad 3.2\text{-}84c$$

which is the same form as equation 3.2.2b in the case of LPC vocoders. The transfer function of the all pole filter is given by equation 3.2-84d. The all pole filter is usually refered to as the (Linear Predictive Coding) LPC filter:

$$\frac{X(Z)}{D(Z)} = \frac{1}{1 - \sum_{k=1}^{P} a_k Z^{-k}} \qquad 3.2\text{-}84d$$

and its spectrum is given by 3.2-83

Note that the inverse filter

$$\frac{D(Z)}{X(Z)} = A(Z) = 1 - \sum_{k=1}^{P} a_k Z^{-k} \qquad 3.2.84e$$

has the power spectrum that is the inverse of the power spectrum in 3.2-83. The power spectrum of the inverse filter can be easily computed since its impulse response from 3.2-84e is given by the sequence

$$1,\ -a_1,\ -a_2, \ldots, -a_p, 0, 0, 0 \ldots \qquad 3.2\text{-}85$$

i.e. it has only P+1 nonzero terms and therefore a short time FFT can be performed on the impulse response given by 3.2-85 without any errors arizing from the truncation of terms that would have been necessary if the (infinite) impulse response of the all pole filter was used for the computation. Any number of zeros can be appended to the sequence 3.2-85 to obtain a desired frequency resolution. Typical amplitude responses of the all pole filter are shown in figures 3.2-23 and figure 3.2-24, for a voiced and unvoiced sound respectively. The time segments corresponding to these spectra are shown as inserts. Also shown are the short time spectra of these speech segments. The correct level match can be obtained by noting that

$$|X(W)|^2 = \frac{|D(W)|^2}{\left|1 - \sum_{k=1}^{P} a_k e^{-jkw}\right|^2} \qquad 3.2\text{-}86$$

and since D(W) is assumed to be flat for a large prediction order P the level matched filter is

$$\frac{\sigma_d^2}{|1 - \sum_{k=1}^{P} a_k e^{-jkw}|^2} \qquad 3.2\text{-}87$$

note that

$$\frac{\sigma_d}{1 - \sum_{i=1}^{P} a_k Z^{-k}} \qquad 3.2\text{-}88$$

corresponds to the filter H(Z) used as a vocal tract model. (eq. 3.2-1 and 3.1-2).

From figures 3.2-23 and 3.2-24 can be seen that the filter models the peaks of the short-time speech spectrum better than the valleys. This is a feature of the LPC filter, related to the fact that it is an all pole filter.

3.2.3.2e LPC Quantization

Quantization of the prediction coefficients a_k directly can easily lead to instabilities in the resulting filters. R. Viswanathan and J. Makhoul [50] undertook a study to find the "best" parameters for representing an all pole filter of the form given in equation 3.2-88, where "best" is related to quantization performance subject to some predetermined criterion. The criterion used was based upon the assumptions that (a) the quality of the synthesized speech is a function f the "maximum perceptual error" between the synthesized and original speech, and (b) that an accurate representation of the power spectrum is necessary for synthesized and original speech. Their criterion for optimal quantization was therefore to minimize the maximum spectral error due to quantization. This is different form the criterion of minimizing σ_d^2 in equation 3.2-51 and their quantization criterion is more suited to LPC vocoder designs. Under certain conditions though, the two criteria can be considered as equivalent [51].

In measuring spectral sensitivity the PARCOR coefficients k_i were used as the independent variable. These can be shown to be the same quantities as the reflection coefficients r_i as defined by equation 3.1-1 [6] and are therefore related to corresponding area coefficients defined by

$$A_i = A_{i+1} \frac{1 + k_i}{1 - k_i}, \quad A_{p+1} = 1, \quad 1 \leqslant i \leqslant p \qquad 3.2\text{-}89$$

The log spectrum S was used to represent the model and a spectral sensitivity was defined as

$$\frac{\partial S}{\partial k_i} = \lim_{\Delta k_i \to 0} \left| \frac{\Delta S}{\Delta k_i} \right| \qquad 3.2\text{-}90a$$

i.e.

$$\frac{\partial S}{\partial k_i} = \lim_{\Delta k_i \to 0} \left| \frac{1}{\Delta k_i} \left[\frac{1}{2\pi} \int_{-\pi}^{+\pi} \right| \log P(k_i, w) - \log P(k_i + \Delta k_i, w) \, dw \right] \right| \qquad 3.2\text{-}90b$$

where P(.,w) is defined by equation 3.2-83.

Typical spectral sensitivities $10\log_{10}(\partial S/\partial k_i)$ are shown in figure 3.2-25. From this figure one can deduce that for a flat spectral sensitivity a nonlinear quantizer need be used for quantizing the reflection coefficients. Flat spectral sensitivity is an equivalent criterion to minimizing the maximum spectral error through optimal quantization. This is similar to the situation where minimizing the prediction error σ_d^2 leads *to a flat prediction error spectrum. Another similar* situation is minimizing reconstruction error in frequency domain coders which also results in a flat error spectrum.

A suitable transformation was found to be

$$f(k_i) = \log \frac{1 + k_i}{1 - k_i} \qquad 3.2\text{-}91a$$

which can also be written as

$$f(k_i) = \log \frac{A_i}{A_{i+1}} \qquad 3.2\text{-}91b$$

by virtue of equation 3.2-89.

The quantities $f(k_i)$ are called log area ratios and provide an approximately optimal set of coefficients for quantization. Equivalent spectral sensitivities $10\log_{10}$ $(\partial S/ \partial f(k_i)$ are shown in figure 3.2-26.

Note that quantization of $f(k_i)$ always leads to a stable filter since the region $-\infty<f(k_i)<+\infty$ always maps into $-1\leq k_i<1$.

3.2.3.2.e1 <u>Optimum bit allocation</u>

Assume that the P parameters for quantization are q_1, $q_2 \ldots q_P$ each allocated bits M_1, $M_2 \ldots M_P$ respectively to a total number of M. The number of levels N is given by

$$N = 2^M \qquad 3.2\text{-}92$$

The quantization step size for q_i is given by

$$\delta_i = \frac{\bar{q}_i - \underline{q}_i}{N_i} \qquad 3.2\text{-}93$$

where $\bar{q}_i$ and $\underline{q}_i$ are the upper and lower bounds on q_i respectively.

The total spectral deviation ΔS defined as

$$\Delta S = \sum_{i=1}^{P} \left| \frac{\partial S}{\partial q_i} \Delta q_i \right| \qquad 3.2\text{-}94$$

is minimized for a constant step size given by

$$\delta = \left[\frac{\prod_{i=1}^{P} (q_i - q_i)}{2^M} \right]^{1/P} \qquad 3.2\text{-}95$$

The appropriate number of levels for each coefficient can then be calculated from 3.2-93 and the required number of bits from 3.2-92. An optimum scalar quantization scheme can therefore be applied by using the log area ratios defined by equation 3.2-91 as the parameters q_i to be quantized. Further reduction in the bit rate required to quantize the all-pole filter parameters can be achieved by taking into account of the fact that linear and nonlinear dependencies exist amongst the parameters to be quantized. One quantization method that takes advantage of the above dependencies is vector quantization [52].

3.2.3.3 Vector Quantization

We have seen that an optimum quantizer adjusts its step size according to the Pdf of the course, providing a smaller step size through the region where the Pdf is high and a larger step size where the Pdf is low. An optimum quantizer can also be designed for the case where the source is multidimensional i.e. when each successive source output X(n) is no longer a scalar but a vector quantity $X(n) = \{X_1(n), X_2(n), X_3(n) \ldots X_K(n)\}$. A scalar quantizer assigns each input X(n) to an output value Y(n). A vector quantizer performs an analogous operation by assigning a vector $Y(n) = \{Y_1(n), Y_2(n), Y_3(n) \ldots Y_K(n)\}$ to each input vector X(n). As in the case of a scalar quantizer there is only a finite number N of reconstruction code vectors Y(n) forming an alphabet or codebook A. This is analogous to the number of reconstruction levels N of the scalar quantizer, and the number of bits required to identify each vector Y(n) is given by

$$M = \log_2 N \qquad 3.2\text{-}96$$

The determination of the optimum reconstruction vectors in A depends on the joint probability distribution function $P(X_1, X_2, X_3 \ldots X_n)$ and an appropriate distortion criterion. In the case of the scalar quantizer the criterion used was the noise variance σ_q^2 and the design of the optimum quantizer minimizes

$$\sigma_q^2 = \int_{-\infty}^{\infty} [y-x]^2 P_x(x)\, dx \qquad 3.2\text{-}97$$

(equation 3.2-9c)

If now the criterion is some other function d(X,Y) of X and Y the formula above can be written as

$$\text{minimize } D(x,y) = \int_{-\infty}^{+\infty} d(x,y)P(x)\,dx \qquad 3.2\text{-}97$$

for the m.s.e. case

$$d(x,y) = [y - x]^2 \qquad 3.2\text{-}98$$

by direct analogy, for the vector case, we need to minimize:

$$D(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} d(X,Y)P(X_1,X_2,\ldots X_k)\,dx_1dx_2\ldots dx_k \qquad 3.2\text{-}99$$

The vectors X and Y can be represented by points in an K-dimensional space. High P(X) = $P(X_1, X_2, X_3,\ldots,X_K)$ implies a dense region in the space whereas a low P(X) implies a sparse region in the space. Optimal quantization will place more reconstruction vectors Y in the dense region(s) of the space whereas fewer (or none) vectors will be placed in the sparce region(s) of the space. Assuming that the alphabet A is known i.e. the positions of the Y vectors are known in the space, the quantizer Q(X) assigns vector Y_i to vector X (i.e. quantizes vector X to the value Y) according to the nearest neighbour rule:

$$Q(X) = Y_i \quad \text{iff} \quad d(X,Y_i) \leq d\,(X,Y_j)$$

$$j \neq i \quad 1 \leq j \leq N \qquad 3.2\text{-}100$$

Note that d(X,Y) in equation 3.2-100 need not be the same as d(X,Y) in equation 3.2-99, although, if one distortion measure provides an optimal criterion the other may not. In particular if the distortion measure in 3.2-100 is replaced by a function F[d(X,Y)] which is strictly monotonic with d(X,Y) i.e.

$$Q'(X) = Y'_i \quad \text{iff} \quad F[d(X,Y'_i)] \leqslant F[d(X,Y'_j)]$$

$$j \neq i \quad 1 \leqslant j \leqslant N \qquad 3.2\text{-}101$$

then exactly the same quantization vector Y_i will be chosen in both cases i.e. $Q'(X) = Q(X)$ and $Y_i = Y'_i$. The same cannot be said for equation 3.2-99 where different alphabet A will result, in general, if $d(X,Y)$ is replaced by $F[d(X,Y)]$.

Equation 3.2-100 also defines a cell C_i known as nearest neighbour cell, voronoi cell or Dirichlet region for each Y_i containing all the points X satisfying 3.2-100. For the purpose of quantization all points (vectors) satisfying 3.2-100 belong to the cell C_i. This condition also defines the vector Y_i if the cell C_i is known: To minimize $D(X,Y)$ in equation 3.2-99 one has to minimize

$$D_i(X,Y_i) = \int_{X \in C_i} d(X,Y_i)\ P(X)\ dX \qquad 3.2\text{-}102$$

for all cells C_i

where

$$P(X)dX = P(X_1,\ X_2, \ldots X_k)dX_1 dX_2 \ldots dX_k \qquad 3.2\text{-}103$$

since

$$D(X,Y) = \sum_{i=1}^{N} D_i(X,Y_i) \qquad 3.2\text{-}104$$

The vector Y_i minimizing 3.2-102 is called the centroid of the cell C_i.

$$Y_i = \text{cent}\ (C_i) \qquad 3.2\text{-}105$$

3.2.3.3a Codebook generations - The LBG algorithm

Note that knowing the cell C_i the vector Y_i can be found from 3.2-102 and conversely, knowing the vector Y_i, the cell C_i can be found from equation 3.2-100.This leads to an iterative procedure for determining the Alphabet A from the distribution $P(X) = P(X_1, X_2.\ldots X_k)$. In practice P(X) is not known. Instead a long training sequence of data is available $\{X(n), 1\leq n\leq L\}$. Using this data the following algorithm can be used to obtain the alphabet A:

Step:1 Initialization: Set m=0. Chose an initial estimate of code vectors $Y_i(0)$ $1\leq i\leq N$ of alphabet A(0)

Step:2 Classification: Classify the set of training vectors $\{X(n), 1\leq n\leq L\}$ into the clusters (cells) C_i according to 3.2-100

Step: 3 Code vector Updating: $m\rightarrow m+1$. Update the code vector of every cluster by computing the centroid of the training vectors in each cluster. $Y_i(m) = \text{cent}\ (C_i(m))$ $1\leq i\leq N$ according to 3.2-102 (3.2-106)

Step: 4 Termination test: If the decrease in the overall distortion D(m) at iteration m relative to D(m-1) is below a certain threshold, stop; otherwise go to step 2.

The above is the k-means algorithm [52].

Equation 3.2-102 has to be modified for the case of using $\{X(n), 1\leq n\leq L\}$ instead of p(X) for its calculation. Then

$$D_i(X,Y_i) = \frac{1}{L_i}\Sigma\ d(X,Y_i) \qquad 3.2\text{-}106$$

$$X \epsilon C_i$$

D(m) in step 4 is given by 3.2-104 as D(X, Y(m))

Test 4 signals the exit from the algorithm if

$$\frac{D(m+1) - D(m)}{D(m)} < E \qquad 3.2\text{-}107$$

where B is a small value, say 0.1%. Any other reasonable termination test may be used in the algorithm. Also a maximum limit on m can be placed.

One method to provide the initial alphabet A(0) for step 1 also results in obtaining designs for quantizers with $N = 2^M$, M = 0,1,... until an initial guess for the N_{final} level quantizer required is obtained. This is as follows

1. Initialization: Set N = 1 and define Y_1 = cent ({X(n), 1≤n≤L}) the centroid of the entire training data.

2. Given a codebook A(N) containing N vectors {Y_i; i=1,...N} "split" each vector Y_i into two close vectors $Y_i + Z$ and $Y_i - Z$ where Z is a fixed perturbation vector. The new codebook has 2N vectors. Replace N by 2N.

3. Is $N = N_{final}$? if so set $A(0) = A(N_{final})$ and exit A(0) is then the initial reproduction alphabet for the N-level quantization algorithm. If $N \neq N_{final}$ run the k-means algorithm using the initial estimate provided by step 2 to obtain a new codebook A(N) and return to step 2.

The k-means algorithm together with the splitting procedure described above is referred to as the LBG algorithm [53, 54]. If provides a locally optimum codebook.

3.2.3.3b Applications - nearest neighbour measures

Since the log area ratios LAR are known to have good quantization properties one way to apply vector quantization to LPC is to quantize the LAR using a m.s.e. measure, i.e. replace the components of the vector X by the LARs and use the Euclidean distance between source and reconstruction vectors as the distortion measure. For a squared error criterion the centroid of the cell is the Euclidean center of gravity given by

$$Y_i = \text{cent}\ (C_i) = \frac{1}{L_i} \sum_{X(n) \epsilon C_i} X(n) \qquad 3.2\text{-}108$$

where L_i is the number of vectors X(n) in cell C_i provided

$$d(X,Y_i) = (X-Y_i)\ (X-Y_i)^t \qquad 3.2\text{-}109$$

In general a mean square criterion will attempt to make the distortion equal in each component of the input vector i.e. the average distortion in each vector component will be approximately equal. In the case where the vector components have widely different variances the components with the smallest variances will have negative SNRs for certain low bit rates in spite, in fact, because of the fact that the vector quantizer is an optimum quantizer. The source vector could therefore be "truncated" in some cases to those components that will have positive SNRs of some prefered value for a particular bit rate. This is true for the LARs since a "truncated" vector will simply represent a filter of lower order (provided the smallest LARs are those of highest indices). The bit rate of the vector quantizer can therefore be used to decide on the useful prediction order than can be accommodated by the quantizer.

Different error distributions can be achieved using various transformations on the vector components. For example if instead of $X(n) = \{X_1(n), X_2(n), X_3(n)...X_k(n)\}$ the transformed vector $X'(n)$ is used defined by

$$X'(n) = \{\log X_1(n), \log X_2(n), ..., \log X_k(n)\} \qquad 3.2\text{-}110$$

under a square error distortion measure given by 3.2-109 with X' replacing X, the percentage error in each component will be approximately constant.

This can be demonstrated by taking two of the components and assuming constant average distortion in each component:

$$Y'_1 = X'_1 + d \qquad 3.2\text{-}111a$$

$$Y'_2 = X'_2 + d \qquad 3.2\text{-}111b$$

or

$$Y'_1 = \log X_1 + d \qquad 3.2\text{-}111c$$

$$Y'_2 = \log X_2 + d \qquad 3.2\text{-}111d$$

to obtain the quantized version of X the inverse transforms can be operated upon Y' to give

$$Y_1 = EXP[Y'_1] = aX_1 \qquad 3.2\text{-}112a$$

$$Y_2 = EXP[Y'_2] = aX_2 \qquad 3.2\text{-}112b$$

where

$$a = EXP[d] \qquad 3.2\text{-}112c$$

in practice of course the equations 3.2-111, 3.2-112 will hold for statistical averages.

Therefore although Vector Quantization will produce an "optimum" solution careful choice of the representation vector should be made for a useful quantizer. This of course can be incorporated into the choice of a meaningful distortion criterion or distance function.

We have seen that the aim of linear predictor designs is to minimize the variance σ_d^2 of the prediction error d(n) [eq. 3.2-50 and 3.2-51]. We will now develop a distortion criterion that can be used for vector quantization of the filters which also aims to minimize the variance of the prediction error.

Assume that we have a windowed sequence of data X(n) where X(n) = 0 for n<0 and n>N-1. The prediction error is given by

$$d(n) = \sum_{i=0}^{P} a_i X(n-i) \qquad 3.2\text{-}113$$

with a_0 = 1 and the minus sign of 3.2-50 incorporated into the a_i coefficients. The total squared prediction error (residual) energy is given by

$$\alpha = \sum_{n=-\infty}^{\infty} [d(n)]^2 \qquad 3.2\text{-}114$$

as in the case of the autocorrelation method. The coefficients $\{a_i,\ i=1,2...P\}$ have been obtained by minimizing the prediction error energy. This minimum value is given by α above. The filter

$$A(Z) = 1 + \sum_{i=1}^{P} a_i Z^{-i} \qquad 3.2\text{-}115$$

is the filter that minimizes α. Let $A(Z)$ represent the uncoded filter. If $\{X(n)\}$ is passed trough the coded filter $A'(Z)$ given by

$$A'(Z) = 1 + \sum_{i=1}^{P} a'_i \ Z^{-i} \qquad 3.2\text{-}116$$

the residual energy δ must be greater than the minimum residual error i.e.

$$\delta = \sum_{n=-\infty}^{\infty} [\sum_{i=0}^{P} a'_i X(n-i)]^2 \geq \alpha \qquad 3.2\text{-}117$$

with equality holding iff $A(Z) = A'(Z)$.

The ratio δ/α is called the likelihood ratio. [51] Evaluation of these ratios can be efficiently carried out through the use of autocorrelation sequences. Let $\{r_a(n)\}$ and $\{r_x(n)\}$ denote the autocorrelation sequence of the polynomial $A(Z)$ and the data $\{X(n)\}$ respectively. The minimal residual error can be computed from [51, 54].

$$\alpha = \sum_{n=-P}^{P} r_a(n) \ r_x(n) \qquad 3.2\text{-}118a$$

$$= r_a(0) \ r_x(0) + 2 \sum_{n=1}^{P} r_a(n) \ r_x(n) \qquad 3.2\text{-}118b$$

with

$$r_a(n) = \sum_{k=0}^{P-n} a_k a_{k+n} \qquad n=0,1,\ldots P \qquad 3.2\text{-}119$$

and

$$r_x(n) = \sum_{k=0}^{N-1-n} X(k)X(k+n) \qquad n=0,1...P<N \qquad 3.2\text{-}120$$

similarly

$$\delta = \sum_{n=-P}^{P} r'_a \, (n) \, r_x(n) \qquad 3.2\text{-}121$$

Our aim is to minimize δ the prediction error variance when the coded inverse filter is used. Note that α is assumed to have been minimized by definition. One distortion measure between the derived coded filter and each individual training sequence X is therefore

$$d(X,\ A'_i) = \delta \qquad 3.2\text{-}121b$$

where A'_i represents the filter A'(Z) (or 1/A'(Z)).

By virtue of equation 3.2-106 the centroid of a particular cell C_i is that filter A'_i minimizing

$$D_i = (X, A'_i \ ) = \frac{1}{L_i} \sum_{X \epsilon C_i} d(X, A'_i) \qquad 3.2\text{-}122a$$

which can also be expressed in the following forms

$$D_i(X, A'_i) = \frac{1}{L_i} \sum_{X \epsilon C_i} \delta \qquad 3.2\text{-}123b$$

$$D_i(X, A'_i) = \frac{1}{L_i} \sum_{X \epsilon C_i} \sum_{n=-P}^{P} r'_a(n) \, r_x(n) \qquad 3.2\text{-}123c$$

$$D_i(X, A'_i) = \sum_{n=-P}^{P} r'_a(n) \frac{1}{L_i} \sum_{X \epsilon C_i} r_x(n) \qquad 3.2\text{-}123d$$

or

$$D_i(X, A'_i) = \sum_{n=-P}^{P} r'_a(n)\, R_x(n) \qquad 3.2\text{-}123e$$

or

$$R_x(n) = \frac{1}{L_i} \sum_{X \epsilon C_i} r_x(n) \qquad 3.2\text{-}124$$

The centroid can therefore be calculated from the autocorrelation equation 3.2-60 where the autocorrelation terms are given by 3.2-124. We now therefore have a distortion measure (eq. 3.2-121) to perform a nearest neighbour search and a formula to calculate the cell centroids (3.2-60 and 3.2-124). This is sufficient to design a codebook using the LBG algorithm.

The magnitude of δ does not only depend on the accuracy of quantization but also on the individual speech segments X. Therefore speech segments with large α will also have large δ and this will affect the quatization process through equation 3.2-123. Segments with large α will be quantized with greater accuracy and vice versa. This is in a sense undesirable since α is proportional to speech energy from 3.2-113 and 3.2-114. In effect if two otherwise identical sentences in the training sequence have been recorded at different sound levels their corresponding predictors will be quantized differently even though they have identical all-pole filters. One way to overcome this is to use the likelihood ratio δ/α instead of δ as the distortion measure i.e.

$$d(X,\ A'_i) = \delta/\alpha \qquad 3.2\text{-}125a$$

this can also be written as

$$d(A_i,\ A'_i) = \delta/\alpha \qquad 3.2\text{-}125b$$

which implies that the distortion criterion measures (quantization) distortion between coded and uncoded filters. Note that since α is fixed by definition we are again minimizing the (weighted) prediction error energy from the coded filter: Nearest neighbour calculations with either 3.2-121b or 3.2-125 will yield the same reconstruction vector A'_i. Equation 3.2-123 now becomes

$$D_i(A_i A'_i) = D_i(X, A'_i) = \sum_{n=-P}^{P} r'_a(n) \frac{1}{L_i} \sum_{X \epsilon C_i} \frac{r_x(n)}{\alpha} \tag{3.2-126a}$$

or

$$D_i(A_i, A'_i) \sum_{n=-P}^{P} r'_a(n) R_x(n) \tag{3.2-126b}$$

with

$$R_x(n) = \frac{1}{L_i} \sum_{X \epsilon C_i} \frac{r_x(n)}{\alpha} \tag{3.2-127a}$$

or

$$R_x(n) = \frac{1}{L_i} \sum_{X \epsilon C_i} \frac{r_x(n)}{\sum_{n=-P}^{n=P} r_a(n) r_x(n)} \tag{3.2-127b}$$

Once again the centroid can be calculated from 3.2-60 with the autocorrelation terms given by $R_x(n)$ of 3.2-127. From 3.2-127b can be seen that normalized autocorrelation terms $r_x(n)$ can be used since they appear both in the numerator and denominator of the equation. Although the formulations above depend heavily on the autocorrelation method, the terms $r_x(n)$ can also be derived from 3.2-60 (This is the autocorrelation matching property [5, 54]) if the coefficients {a_i i=1,2....P} have been derived through some other method e.g. the covariance or Burg method. One then proceeds as above to quantize the filters without the need to calculate $r_x(n)$ from the speech segments directly. The above formulation using 3.2-125 as the distortion measure is equivalent to the gain separated Itakura-Saito distortion measure [54]. It is reported that ([52]) Vector Quantization (VQ) employing forms of the Itakura-Saito measure provides similar perceptual performance as the squared error distortion measure on the LARs for full band signals (i.e. sampling rate ~ 8KHz, P~12).

3.2.3.3c Other issues in Vector Quantization

A major concern in VQ is algorithm complexity: At the encoder, the source vector must be compared with every codeword in the codebook to find its nearest neighbour in order to code the source vector with the minimum distortion. This involves N comparisons where N is the total number of codewords. Alternatively this involves N calculations of the appropriate distortion measure, and for complex distortion measures and large N the complexity can be prohibitively high. Another factor adding to complexity is the vector size since this will also increase the complexity of the distortion measure calculation.

A second issue is storage: At both the encoder and decoder the N-member codebooks have to be stored in memory.

One way to reduce the complexity is through the binary tree search procedure [52, [54, 55]. In this procedure the N dimensional space is first divided into two regions (using the k-means algorithm with k=2), then each of the two regions is divided into two subregions and so on until the space is divided into L regions.

Associated with each region at each binary division is its centroid. Fig. 3.2-27 is a schematic of binary division of space into L=8 cells. At the first binary division V_1 and V_2 are the region centroids. At the second binary division there are four regions with centroids V_3 through V_6. The centroids of the regions after the third binary division are the code vectors Y_i. An input vector X is quantized by following a path along the tree that gives the minimum distortion at each node in the path: X is compared to V_1 and V_2. If $d(X, V_2)<d(X,V_1)$ for example then the path leading to V_2 is taken. The vectors stemming from V_1 are no longer considered for the search, and so on through the remaining nodes. The total number of distortion computations is now equal to $2\log_2 L$. Non uniform trees can also be constructed by monitoring the occupancy of each region. This is shown in figure 3.2-28. A performance comparison between scalar, binary and full search quantization is shown in figure 3.2-29. The spectral error reflects the rms log spectral deviation of equation 3.2-90. It can be seen that the performance of the binary search is only slightly inferior to the full search compared to the scalar case.

Another technique aimed to reduce storage and computational cost is cascaded quantization. (Figure 3.2-30). The input vector X is first quantized using a B_1 bit (L_1-level) VQ. The residual E between X and its quantized value Z_i is then used as the input to a B_2-bit (L_2-level) second VQ stage with output W_j.

The final quantized value is the sum of the two vectors Z_i and W_j. The performance of this quantizer is inferior to the conventional VQ since a lot of the dependencies between the vectors are lost through the operation. The KLT matrix A serves to reduce this loss [52] but can only improve on linear dependencies.

Another method of reducing complexity and storage is through the use of split-band codes: The spectral envelope is split into a number of bands and each band is vector quantized separately [56]. We have seen that this technique also reduces computation for the derivation of the uncoded filters.

Other variations include product codes [57, 58], use of transforms prior to quantization [59, 60], Stochastic codebooks [61, 62], Adaptive Vector Quantization [63], Predicive vector Quantization [64] and Segment Quantization [65].

We have seen how, given a set of training data $\{X(n), 1 \leq n \leq L\}$, a distortion measure and a means to calculate centroids one can construct a codebook for a vector Quantizer. We will now consider several issues concerning the training data.

Ideally the training set of data should contain every possible vector that is likely to be produced by the source and, additionally, to provide a sample pdf that is identical to the pdf of the source. The last feature implies that the relative occurrence of a vector in the training set should match the relative occurrence form the source, or alternatively, the training set should have the same density in the multidimensional space as the source. For speech signals this implies that both the framesize and the frame rate of the training set should be the same as the one intended to be used in the coding. Note that preferential weighting to certain types of spectra can be given by increasing their number in the training set at the expense of the other spectral types. This can also be achieved by e.g. increasing the frame rate: longer duration sounds such as vowels will then be preferentially encoded at the expense of shorter or transitional segments. The correct choice can only be decided through perceptual studies. The very first feature mentioned above cannot be achieved in practice since it would imply a training set incorporating every possible sound from every possible speaker which is clearly impractical. One therefore is forced to use a limited amount of training data. 50 training data per coding vector are considered sufficient although similar performance can probably be achieved with a few as 20 training vectors/codeword whereas 10 training vectors/codeword will probably result in a noticeable but perhaps acceptable deterioration. Figure 3.2-31 shows the mse as a function of the number of training vectors per level (per codeword). Clearly this is only an example and the necessary number of training vectors must be determined from a similar plot such as the above. The figure also shows the performance of the codebook on data that were not used for training. Note that the two curves are different, exactly because not all possible training vectors can be used. The two plots follow a parallel course for high numbers of training vectors because the independent test data were produced by different speakers. The difference is related to the Robustness of the codebook: The codebook can have a good performance with certain speakers and bad for others. Robustness can be improved by including as many speakers as possible in the training data.

3.2.3.4 Backward Predictor Adaptation

The coding of the LPC parameters helped to introduce the concept of vector quantization resulting in (locally) optimum quantizers for the all-pole filters. If on the other hand, predictor coefficients can be estimated on the basis of quantized and transmitted data there is no need for the use of a filter quantizer. This kind of predictor is called a sequentially adaptive predictor. Since no side information is required the predictor can be updated as often as desired, usually from sample to sample. Most algorithms are based on the method of steepest descent or gradient search [21]. The general structure is shown in figure 3.2-15b. For high bit rates APF and APB provide comparable performance. For lower bit rates, APF performs better since APB gains are limited by quantization effects: All pole predictors of high orders are more susceptible to quantization and transmission noise and therefore predictors of small order prevail for robust APB algorithms.

3.2.3.5 Spectral Fine Structure Predictors

We have seen that envelope predictors exploit redundancies in the speech signal that stem from the formant structure of the signal. For voiced sounds further redundancies can be exploited, those stemming from the signal's fine or pitch structure. The corresponding predictors are called long delay predictors because distant samples are utilized to obtain the difference (prediction error) signal, as opposed to envelope predictors which are also called short delay predictors, involving only near samples. The two predictors can be used in either order but, usually, envelope prediction is performed first and, then, pitch prediction is performed on the prediction error signal from the first stage. Coders employing both short and long delay predictors are usually refered to as APC or Adaptive Predictive Coders.

The long delay synthesis filter is of the form

$$\frac{1}{B(Z)} = \frac{1}{1-P(Z)} \qquad 3.2\text{-}128a$$

with the predictor P(Z) being usually of third order

$$P(Z) = b_1 Z^{-M+1} + b_2 Z^{-M} + b_3 Z^{-M-1} \qquad 3.2{=}128b$$

or first order,

$$b_1 = b_3 = 0, \quad P(Z) = b_2 Z^{-M} \qquad 3.2\text{-}128c$$

The corresponding analysis filter is of course

$$B(Z) = 1 - P(Z) \qquad 3.2\text{-}128d$$

operating on the (first) prediction error obtained from envelope prediction, d(n) given by equation 3.2-50 to produce the second prediction error e(n) given by

$$e(n) = d(n) - \sum_{K=1}^{3} b_k d(n-M+2-K) \qquad 3.2\text{-}129$$

The optimum predictor is obtained, as in the case of short-term prediction, by minimizing the energy of e(n) over the time interval of interest i.e.

$$\text{minimize } E = \sum_{PL} [e(n)]^2 \qquad 3.2\text{-}130a$$

PL the minimization range is usually the whole current analysis frame as in the covariance method for the short term predictor:

$$\text{minimize } E = \sum_{n=o}^{N-1} [e(n)]^2 \qquad 3.2\text{-}130b$$

by virtue of equation 3.2-129 and using the one tap predictor from 3.2-128c the above equation can be written as

$$\text{minimize } E = \sum_{n=0}^{N-1} [d(n)-b_2 d(n-M)]^2$$

which can be solved by setting

$$\partial E/\partial b_2 = 0 \qquad 3.2\text{-}131$$

$$\frac{\partial E}{\partial b_2} = - \sum_{n=0}^{N-1} 2[d(n)-b_2 d(n-M)]d(n-M) = 0 \qquad 3.2\text{-}132a$$

or

$$b_2 = \frac{\sum_{n=0}^{N-1} d(n)d(n-M)}{\sum_{n=0}^{N-1} d^2(n-M)} \qquad 3.2\text{-}132b$$

substituting this value for b_2 into equation 3.2-131 gives for the minimum E:

$$E_{min} = \sum_N [d^2(n)] - \frac{[\sum_N d(n)d(n-M)]^2}{\sum_N d^2(n-M)} \qquad 3.2\text{-}133$$

where the summation over the N values of n is as before. It can be seen from 3.2-133 that the minimum value of E_{min} is obtained for that M which maximizes

$$DE(M) = \frac{[\sum_N d(n)d(n-M)]^2}{\sum_N d^2\ (n-M)} \qquad 3.2\text{-}134$$

The function DE(M) is calculated for an expected range of M values, for example, an equivalent delay of 2.5msec to 18.5msec which, for 8KHz sampled speech gives a range of M between 20 and 147. The value of M can then be coded using 7 bits and covers the expected time length for the pitch period of most speakers. Note that the value of M will frequently require past samples d(n-M) outside the current analysis frame for the calculation of DE(M). Once the value of M is found, b_2 can be calculated from the 3.2-132b.

For the case of the three tap predictor equation 3.2-130 can be expanded into

$$\text{minimize } E = \sum_{n=0}^{N-1} [d(n) - \sum_{k=1}^{3} b_k d(n-M+2-K)]^2 \qquad 3.2\text{-}135$$

which can be solved by setting

$$\frac{\partial E}{\partial b_1} = \frac{\partial E}{\partial b_2} = \frac{\partial E}{\partial b_3} = 0 \qquad 3.2\text{-}136a$$

thus obtaining three simultaneous equations for the minimization problem.

$$b_1X_1 + b_2 X_2 + b_3X_3 = D_1 \qquad 3.2\text{-}136b$$

$$b_1X_2 + b_2X_4 + b_3X_5 = D_2 \qquad 3.2\text{-}136c$$

$$b_1X_3 + b_2X_5 + b_3X_6 = D_3 \qquad 3.2\text{-}136d$$

where

$$X_1 = \sum_N [d(n-M+1)]^2 \qquad 3.2\text{-}136e$$

$$X_2 = \sum_N [d(n-M+1)d(n-M)] \qquad 3.2\text{-}136f$$

$$X_3 = \sum_N [d(n-M+1)d(n-M-1)] \qquad 3.2\text{-}136g$$

$$X_4 = \sum_N [d(n-M)]^2 \qquad 3.2\text{-}136h$$

$$X_5 = \sum_N [d(n-M)d(n-M-1)] \qquad 3.2\text{-}136i$$

$$X_6 = \sum_N [d(n-M-1)]^2 \qquad 3.2\text{-}136j$$

$$D_1 = \sum_N d(n)d(n-M+1) \qquad 3.2\text{-}136k$$

$$D_2 = \sum_N d(n)d(n-M) \qquad 3.2\text{-}136l$$

$$D_3 = \sum_N d(n)d(n-M-1) \qquad 3.2\text{-}136m$$

The value of M can be obtained by maximizing 3.1-134, although this is only an approximation. The prediction errors resulting from envelope and pitch prediction can be seen in figure 3.2-32. The second prediction error is nearly Gaussian (figure 3.2-34) and looks like white noise during steady speech segments.

The two analysis (and synthesis) filters can be put schematically in cascade forms. One then is tempted to provide a composite response from

$$\frac{1}{A(Z)} \cdot \frac{1}{B(Z)} = \frac{1}{1-\sum_{k=1}^{P} a_k Z^{-k}} \cdot \frac{1}{1-\sum_{k=1}^{3} b_k Z^{-M+2-k}}$$

$$= \frac{1}{1-\sum_{k=1}^{P} a_k Z^{-k} - \sum_{k=1}^{3} b_k Z^{-M+2-k} + \sum_{k=1}^{P} \sum_{j=1}^{3} a_k b_j Z^{-k-M+2-j}}$$

3.2-137

$$= \frac{1}{C(Z)} \qquad 3.2\text{-}137$$

and use the corresponding filter 1/C(Z) and C(Z) for the synthesis and analysis parts of the algorithms respectively. The above procedure would be quite wrong since the predictor filters are not time-invariant and the Z-transforms given above are only short-time approximations to the infinite-time transforms for which they are defined. The approximations hold well when the impulse response of the filter lasts only over time intervals during which the filter response does not change appreciably. This is true usually for the short-time predictor but not for the pitch predictor. The impulse response of the pitch predictor typically lasts over several pitch periods for voiced sounds.

3.2.3.6 Noise Feedback coding-Noise shaping

In the previous sections it was shown how a DPCM coder can be designed to encode a speech signal in an optimal way. The criterion of optimality was the SNR based on a m.s.e. criterion. A generalized DPCM coder will now be presented. This coder results in a higher m.s.e. than the equivalent DPCM coder, but the perceived quality is better due to noise shaping and resulting noise masking.

From equations 3.2-30 and 3.2-48 the prediction error in DPCM is

$$d(n) = \hat{X}(n) - X(n) = X(n) - \sum_{j=1}^{P} a_j Y(n-j)$$

$$= X(n) - \sum_{j=1}^{P} \cdot a_j X(n-j) + \sum_{j=1}^{P} a_j q(n-j) \qquad 3.2\text{-}138$$

where the $\{a_j \ j=1,2...P\}$ are the coefficients of an all pole filter H(z) defined by

$$H(Z) = \frac{1}{A(Z)} = \frac{1}{1-P(Z)} = \frac{1}{1-\sum_{j=1}^{P} a_j Z^{-j}} \qquad 3.2\text{-}139$$

With A(Z) the corresponding inverse filter and P(Z) the prediction filter.

The quantizer input (prediction error d(n)) can be split into two parts, d*(n) which is the prediction error with zero quantization noise (eq. 3.2-50) and the quantization error feedback term QEF(n) (eq. 3.2-49). Schematically, this is shown in figure 3.2-34a.; As we have seen from eq. 3.2-36, the reconstruction error r(n) is identical to the quantizer error q(n)

$$r(n) = q(n); \quad R(Z) = Q(Z) \qquad 3.2\text{-}140$$

where R and Q are the Z-transforms of r and q respectively. If the quantizer in DPCM is taken out and placed after the prediction loop the scheme in figure 3.2-34b is obtained. For this scheme it is easy to show using difference equations that

$$R(Z) = Q(Z) \cdot \frac{1}{1-P(Z)} \qquad 3.2\text{-}141$$

Figure 3.2-34c shows a generalized predictive coder with a error feedback filter F(Z). This scheme is called noise feedback coding NFC. [35, 36]. For this system:

$$R(Z) = Q(Z) \quad \frac{1-F(Z)}{1-P(Z)} \qquad 3.2\text{-}142$$

[21, 35].

Clearly DPCM and D*PCM are special cases of 3.2-142 with F(Z) = P(Z) and F(Z) = 0 respectively. Assuming a white quantizer error spectrum the reconstruction error spectrum for the three coders is as shown in fig. 3.2-35 (dashed lines) for a particular speech segment, having the all pole spectrum shown in solid lines. Under the assumptions of white quantization error the spectrum of the reconstruction noise is determined only by the factor [1-F(Z)]/[1-P(Z)] as shown in 3.2-142. Let the squared magnitude of this factor at a frequency f be M(f) then

$$M(f) = |[1-F(e^{2\Pi jfT})]/[1-P(e^{2\Pi jfT})]|^2$$

(T is the sampling interval) 3.2-143

equation 3.2-143 implies [35]

$$\frac{1}{f_s} \int_0^{f_s} \log M(f) df = 0 \qquad 3.2\text{-}144$$

i.e. the average value of the log power spectrum of the reconstruction noise is determined solely by the quantizer and is not altered by the filter F or the predictor P. The filter F can be chosen to minimize an error measure in which the noise is weighted according to some subjectively meaningful criterion [67]. A fairly general approach is to minimize the noise power, weighted at each frequency by a function W(f).

From equations 3.2-142 and 3.2-143 the reconstruction power spectral density is given by

$$S_r(f) = S_q(f)M(f) \qquad 3.2\text{-}145a$$

and assuming white quantizer noise

$$S_r(f) = \sigma_q^2 M(f) \qquad 3.2\text{-}145b$$

one could then aim to minimize

$$E = \int_0^{f_s} W(f)\ \sigma_q^2 M(f) = \sigma_q^2 \int_0^{f_s} W(f)M(f) \qquad 3.2\text{-}146$$

under the constraint of eq. 3.2-144. The minimum is achieved for [67, 35]

$$\log S_r(f) = -\log W(f)+\log\sigma_q^2+\frac{1}{f_s}\int_0^{f_s}\log W(f)df$$

3.2-147a

or

$$\log M(f) = -\log W(f)+\frac{1}{f_s}\int_0^{f_s}\log W(f)df \qquad 3.2\text{-}147b$$

or

$$\log|1-F(f)|^2=\log|1-P(f)|^2-\log W(f)+\frac{1}{f_s}\int_0^{f_s}\log W(f)df$$

3.2-147c

therefore if the perceptually derived function W(f) is known the filter F can be calculated from 3.2-147c since its psd is known (by transforming $|1-F(f)|^2$ to an autocorrelation function and obtaining filter coefficients through the autocorrelation equation 3.2-60). Such a procedure was undertaken in [67].

Alternatively a choice for F(Z) can be made to obtain a noise spectum intermediate to that of DPCM and D*PCM by letting

$$F(Z^{-1}) = P(aZ^{-1}) \qquad 3.2\text{-}148$$

a is the noise factor.

For a=1 F(Z)=P(Z) whereas for a=0, F=0. Intermediate values of a serve to increase the bandwidths of the zeros of 1-F with respect to the bandwidths of the zeros of 1-P. This approach results in an error spectrum such as the one shown in figure 3.2-35c. A filter such as the one given in 3.2-148 improves the perceptual quality of speech provided the original noise level is adequately low as in figure 3.2-35a. A suitable value for a is around 0.7 [36]. Such a coder as described above, together with a pitch predictor and appropriate quantization procedures for the resulting residual, forms the basis of APC (adaptive predictive coding) [36].

It is difficult to explain the effects of noise shaping from observations on Noise Feedback coders since the quantizer itself introduces some shaping at low bit rates: The quantizer does not introduce white noise into the signal. The noise is correlated with the input signal. Effective procedures to achieve a white quantizer error spectrum will be described under the section on delayed decision coding.

In order to assess the effects of noise shaping, suitably shaped noise can be introduced into the speech signal not through coding but through direct noise injection. Such a study was undertaken by McDermott *et al* [68]. In the above study the power spectral density (psd) of the noise was related to the psd of the speech by the following expression:

$$P_n(w) = P_s{}^a(w).F^b(w).G \qquad 3.2\text{-}149$$

where $P_n(w)$ is the psd of the noise, $P_s(w)$ the psd of the speech, F(w) a fixed weighting function and G a constant specifying the SNR of the output signal. b is a binary variable taking the values 0 and 1. With b=0 all frequencies have equal weight. With b=1 the noise spectrum is weighted according to the width of the Articulation index bands (or, equivalently, according to the width of the critical bands). G is also a binary variable such that the resulting SNR is either 6 or 12 dB. a takes the values 0.0, 0.25, 0.5, 1.0. As the value increases from 0 to 1, the shape of the spectrum changes from a flat distribution to one that corresponds to the speech spectrum. Two speakers were used, one male and one female. For the 6 dB condition perceptual results (studied through multidimentional procedures) revealed that as the shaping increases, the effect of noise is the same for male and female speech: For low correspondence in the spectra the noise is perceived as a hiss, an additive distortion in the background. As the correspondence increases, the distortion manifests itself as a distortion in the speech signal going through the stages of rumble, hoarse and finally burble as the correspondence between speech and noise spectrums becomes complete.

This series of events are easy to explain through our knowledge about the auditory system: When the noise is white, regions of noise, particularly at high frequencies but also in interformant regions are not masked at all by the speech signal and therefore sound as if the speech signal was not there, i.e. as a hissing sound. Moreover, the formant SNR is high enough such that no effective modulation of the formant harmonics takes place to create the sensation of roughness. In other words the components of noise falling through the critical band (CB) centered around a particular formant are not high enough to cause roughness. As the noise power is increased in the formants and decreased in the valleys the speech signal effectively masks the noise, which loses its perceptual attributes completely and cannot be perceived as a sound on its own. This is the crossover situation which the ear perceives as rumble. Meanwhile the power of the noise in the formants is increased. At some stage, sufficient noise power is concentrated around the formant harmonics to cause the sensation of roughness. This concentration of noise power is essential for the perception of the roughness since only the components of the noise falling within a CB around the particular formant will have a contribution to the sensation. Finally, as the shaping is increased further, the effective bandwidth of the noise around the formants is so narrow so as to allow only slow fluctuations of the formant harmonics which are no longer perceived as roughness but are sufficiently slow to be perceived as loudness fluctuations of the harmonics themselves. The loudness fluctuations of the formant harmonics cause fluctuations in the apparent position of the formants which results in the burbling sensation. Our knowledge about the auditory system has therefore enabled us to describe - at least qualitatively - the series of events occurring through noise shaping. It should now be apparent that models of distortion calculation based upon the estimation of the loudness of noise [69] would be totally inadequate to describe the multitude of events described above in their entirety. Such "unidimensional" approaches can be tuned though to operate around a particular noise shaping degree or "factor" of interest. Such an approach will produce a locally optimum distortion measure [69], tuned to a particular coding technique, noiseshaping range and bit rate. This measure will be grossly suboptimal as one deviates from the conditions it was optimized for.

3.2.3.7 Delayed Decision coding

In the previous sections we have seen that the use of encoding delay in estimating predictor and adaptive quantizer parameters resulted in improved performance over instantaneous coders. Further gains can be achieved if the prediction residual obtained from structures mentioned above is not quantized instantaneously. In conventional quantizers the output value is based on an instantaneous decision based only on the current input value (although, through prediction, the current input value may contain information about "adjacent" samples). Such coding schemes are called single-path coders. In multipath search coding schemes on the other hand an input sequence X(n) or vector X is compared with a collection of possible output sequences $Y_K(n)$ or possible output vectors Y_K.

The optimum output sequence is the nearest neighbour sequence subject to some predetermined distortion criterion, for example a mean square error:

$$E_k = (X-Y_k)^T(X-Y_k) \qquad 3.2\text{-}150$$

figure 3.2-36

Delayed decision coders not only provide a closer approximation to the rate distortion bound, they also enable the use of fractional bit rates/sample, in particular bit rates of less than 1 bit/sample. In addition they sometimes provide a stabilizing action with otherwise unstable coder configurations (such as predictive systems with coarse quantizers).

There are three general classes of delayed decision coders: Codebook Coders, Tree Coders and Trellis Coders. The reconstruction sequences Y_K for populating codebooks trees and trellises can be obtained through deterministic, stochastic or iterative means. These can be either stored at both encoder and decoder or generated from appropriate parameters determined at the encoder and transmitted to the decoder [70].

A codebook coder is identical to the vector quantizer described earlier for the quantization of side information. The codewords Y_K can be determined as before through an iterative procedure [71, 72] or selected stochastically from the output of a random process [73, 74]. A more deterministic approach which lends itself to fast implementation is the use of algebraic codes borrowed from binary error-correction theory [75, 76].

In tree and trellis coding the output sequences cannot be chosen independently but possess a particular structure. Figure 3.2-37 shows sequences of length L arranged in the forms of a tree or trellis of depth L. Its branches are populated with reconstruction values. Different sequences therefore have a number of common elements. Each sequence forms a path through the tree or trellis. The information send to the receiver describes how to trace through the tree or trellis. This information is sometimes called the path map. As before, deterministic or stochastic means can be used to populate the tree or trellis. Examples of tree encoding can be found in [36, 77, 78].

Delayed decision coding has been applied very successfully in coders employing an APC structure. Such a general approach is shown in figure 3.2-77. An innovation sequence V_n appropriately scaled by an adaptive gain term σ is fed through the synthesis filters $1-P_d$ and $1-P_s$ to produce an approximation sequence $\hat{S}_n$. The determination of the gain term σ depends upon the particular method used to represent the sequence V_n. The predictor filters P_d and P_s refer to pitch and envelope prediction respectively. Some schemes [79, 80, 81] can operate successfully without pitch prediction. The approximation sequence $\hat{S}_n$ is compared to the corresponding original speech sequence S_n. The comparison involves a weighted m.s.e. criterion where the weighting function is usually related to the noise shaping filters used in APC. Through various means which depend upon the particular method used, an appropriate sequence V_n is selected which results in a "minimum" weighted m.s.e. Information to completely define the selected sequence is then sent to the receiver together with the side information related to the prediction and the gain σ.

An interesting subclass of the above set of delayed decision coders whose structure has just been described is Multipulse - LPC [79]. For this coder the information about the selected inovations sequence is represented by the amplitudes and locations of nonzero samples in the sequence. Such non-zero samples usually represent a small fraction of the overall sequence. The "optimum" amplitudes and locations are determined iteratively in a succession of stages, each of which determines the amplitude and location of one pulse in the innovation sequence. The optimum amplitude is obtained by setting the derivative of the weighted m.s.e. measure mentioned above with respect to the unknown amplitude to zero. The error measure is then a function only of the pulse location. The optimum location is then found by computing the m.s.e. for all possible locations and by locating its minimum. Additional pulses reduce this minimum further and are defined as above. Details of the multipulse algorithm can be found in appendix F. A similar procedure is that of regular-pulse excitation [81]. The above procedures automatically produce quasi-periodic innovations sequences for voiced speech input without the need of a pitch predictor (although pitch predictors have also been applied [82-84, 81] and produce random excitations during unvoiced speech. Such coders provide useful systems around 9.6kbs/sec but at lower bit rates stochastically determined innovations sequences generally yield better quality speech as in the CELP structures [74].

3.2.3.8 Frequency domain Coders

The redundancy removal and corresponding prediction gain that results form (A)DPCM and APC structures can also be realized in the frequency domain. The category of coder algorithms which have been relatively successful in achieving this goal is the class of frequency domain coders. In this class of coders the speech signal is divided into a set of frequency components which are (usually) separately encoded. In this way different frequency bands can be preferentially encoded according to perceptual criteria for each band, and quantizing noise can be contained within bands. Two basic types of frequency domain coders have been proposed and implemented, namely subband coders and transform coders. In the first case the speech spectrum is partitioned into a set of, typically, 8 contiguous bands by means of a filterbank analysis. In the second case a block by block transform analysis is used to decompose the signal into typically 128 frequency components. Both techniques, in effect, attempt to perform some type of short-time spectral analysis of the input signal although, clearly, the spectral resolution in the two methods is different.

3.2.3.8a Subband Coding

Frequency domain coders operate by removing the redundancy in the input signal in the frequency domain [85]. The advantage of frequency domain over time domain redundancy removal is that the number of bits used to encode each band can be variable which can provide any desired form of noise shaping, something that can be realised to some extent by using noise feedback in the time domain prediction.

In a subband coder the speech signal is divided into a number of 2 to 32 subbands by a bank of bandpass filters. Prior to encoding, each band is in effect low pass translated to zero frequency by a modulation process equivalent to single-side band amplitude modulation. [86, 87, 88]. It is then sampled at its Nyquist rate (twice the width of the band) and then encoded using one of any of the speech coders that were developed in the past to encode the full band signal.

This enables a coding for each band that conforms to perceptual criteria specific to that band. On reconstruction, the subband signals are decoded and modulated back to their original positions. Finally they are summed to give a "close" replica of the original signal.

Figure 3.2-39 illustrates a basic block diagram of the subband coder. The coder consists of a bank of M bandpass filters each followed by its own encoder and a multiplexer. The receiver performs the inverse task of demultiplexing, decoding and bandpass filtering. Finally the subband signals are added to produce the full band band approximation to the input signal. Since individual time waveforms are "closely" approximated at each stage the subband coder is a waveform preserving coder. This is different from the channel vocoder where the object of the filter bank is to preserve the short time energy pattern in the frequency domain.

There are basically two kinds of filter bank responces which can be used to perform the subband splitting. Filters that overlap or filters that do not as shown in figures 3.2-40 a & b. The filters in figure 3.2-40b require extremely fast roll-offs and hence increased delay and complexity but offer the possibility of reduced sampling rates. The perceptual penalty is a reverberant quality in the output speech due to the interband frequency gaps which can be made smaller by using faster roll-offs for the filters. The perceptual effect can be made less noticeable at lower bit rates (hence wider gaps) by adaptively tracking the formant frequencies [89].

The subband width can be different for the different bands as in figure 3.2-40b although this implies a higher complexity. The advantage of unequal bands lies in the different perceptual importance of the different frequency regions. With this design the low frequencies can be given more importance than the higher bands. This is desirable from the perceptual point of view as shown by the articulation index function and other perceptual measures. One example of dividing narrow band speech in the 200 to 3200 Hz frequency range into 4 bands using the Articulation index as the criterion is as follows:

Subband No.	Frequency range (Hz)
1	200 - 700
2	700 - 1300
3	1300 - 2020
4	2020 - 3200

Each band contributes approximately 20% to the A.I. corresponding to a word intelligibility of 93% [86].

Smaller width bands implies that the non-flattness of the spectrum in that particular frequency region can be better exploited. Provided the quantization of the parameters is carefully done in each case, the only difference in performance when dividing a subband into a number of smaller bands is given by the ratio of the mean signal variance in that band to the geometric mean of the variances of the smaller bands (see section on subband gain).

For the lower bit rates small gaps can be permitted between bands over and above those dictated by aliasing considerations to conserve bandwidth and bit rate as shown in figure 3.2-40b, although quality suffers as a result.

In recent years an important filter bank design (the QMF) has the characteristic of equal subband widths. This, together with adaptive bit allocation based on the speech power in each band made the use of higher complexity unequal division unnecessary. The reason for this is that the speech power is higher where the speech signal is more important: The lower frequencies are important for the pitch information for voiced sounds and this region has also high power. The other important characteristic of speech namely the speech formants are important to intelligibility and they are, by definition, higher in power than the speech in surrounding frequencies. Thus thousands of years of adaptation of the speech production and speech perception mechanisms has resulted in a speech signal with optimum characteristics for detection in noise. Another advantage of QMF is that no gaps exist between the bands thus eliminating another form of subband specific distortion.

3.2.3.8a1 Integer Band filter banks

An important feature in figure 3.2-41 is that bandpass cutoffs are chosen such that each band can be sampled at twice the corresponding bandwidth rather than at twice the highest frequency of the full band signal. This is possible in the special situation of integer band sampling [90] where the lower cutoff frequency is an integral multiple of the bandwidth. This approach is particularly attractive for hardware implementation since it eliminates the need for modulators.

Figure 3.2-42 shows the sequence of integer band implementation steps. The speech band is partitioned into N subbands by bandpass filters BP1 to BPn. The output of each filter in the transmitter is resampled at a rate of 2fi where fi is the width of the subband and i refers to the ith subband. The decimation implies a repetition of the spectrum as shown in figure 3.2-42b. One of the repetitions will be in the baseband so that the decimation process automatically translates the lower frequency edge of the bandpass signal to zero frequency. At the receiver, the interpolator fills in zero samples between each pair of incoming lowpass samples so that the sampling rates of the decoder outputs are increased to the original sampling rate of the full band signal. Through this operation a harmonic of the signal is bandpass translated to the appropriate initial bandpass region. Prior to addition the signals are again bandpassed through identical filters BP1 to BPn which act a interpolating filters removing the unwanted images of each subband. The explicit modulation process mentioned above is therefore replaced by the discrete time processes of decimation and interpolation. It is assumed that the interpolation process includes an amplitude scaling factor. This maintains the original value of input variance in spite of the zero valued amplitudes introduced in the interpolaiton process.

It can be shown that even-numbered bands get inverted in the processes. This leads to a shuffling of the bands at the end of a tree structured QMF to be seen later.

The integer band constraint is assumed for minimizing subband sampling frequencies and therefore the overall bit rate.

3.2.3.8a2 Quadrature Mirror Filters (QMF)

We have seen that interband gaps and the increased complexity are undesired effects of nonoverlapping filters. These can be aleviated by using the other type of filters that result in — an overlapping subband division. Figure 3.2-40a suggests that aliasing effects can occur. This problem is diminished when one uses QMF (figure 3.2-43b [88, 91]. This figure shows the division of a full band signal into two of equal width by using a constrained pair of lowpass and highpass filters. By repeated subdivisions one can realize a filter bank with the number of bands, M given by a power of two. Other combinations can also be realized by using tree branches of different depth.

The way that the aliasing problem is solved is explained below: The first stage of the tree will be considered since at each branch the operations are identical.

Each of the subband signals $X_1(n)$ and $X_u(n)$ is resampled by a factor 2:1. The reduction is necessary to maintain a minimal overall bit rate in encoding the signals. This reduction of sampling rate (resampling) is the one that introduces the aliasing terms because of the finite filter roll off. With consideration of figure (3.2-43b) the signal energy in the frequency range above the cutoff frequency of the lowpass filter (which coincides with half the new sampling frequency) will be folded down into the low frequency band after resampling. This will appear as aliasing distortion in the signal, in the frequency range covered by the hatched region in the figure. Similarly energy folded up into the region of the highpass filter will appear as aliasing in this band. The steeper the cut off the smaller will be the hatched areas and so will be the aliasing. This was how the aliasing problem was tackled before the QMFs. To bring the sampling rate back to its original value after coding and prior to addition of the subband outputs, zero-valued samples are inserted between the samples. This creates a periodic repetition of the spectrum in the frequency domain which is then filtered out by the lowpass filter $h_1(n)$ in the receiver. This filter interpolates the samples in the time domain and attainuates the images in the frequency domain. In the same way, the signal in the upper band is repeated in frequency and is repetitions attainuated by the filter in that band. If ordinary filters were used the restoration of the full band signal would depend on the degree that the interpolating filters approximated ideal lowpass and highpass signals. For signals processed through the QMFs though, any remaining components of the images are cancelled out by the aliasing terms introduced in the analysis. The cancelation, which is exact in the absence of coding occurs after the addition of the subband signals.

Aliasing terms of the quantization noise are of course not cancelled out since no such terms were present at the time of splitting into the subbands.

The types of filters that are designed around the QM idea are usually FIR designs which are symmetrical for the lower band and antisymmetrical for the higher band i.e.

$$h_l(n) = h_u(n)=0. \quad 0>n\geq N \qquad 3.2\text{-}151$$

$$h_l(n) = h_l(N-1-n) \qquad n=0,1,...N/2-1 \qquad 3.2\text{-}152$$

$$h_u(n)=-h_u(N-1-n) \qquad n=0,1,....N/2-1 \qquad 3.2\text{-}153$$

The filters must also satisfy the condition:

$$h_u(n) = (-1)^n h_l(n) \qquad n=0,1,...N-1 \qquad 3.2\text{-}154$$

which describes the mirror image relationship of the filters as shown in figure 3.2-44. From the above it is obvious that only (N)/2 filter coefficients need be stored for both filters. The combined filter response must also exhibit an all pass characteristic i.e.

$$|H_l(W)|^2 + |H_u(W)|^2 = 1 \qquad 3.2\text{-}155$$

where $H_l(W)$ and $H_u(W)$are the fourier transforms of hl(n) and hu(n) respectively. The all pass characteristic cannot be met exactly but can be closely approximated for modest values of N. To obtain the filter coefficient an optimization program can be used, alternatively already tabulated values can be used [92]. An example of the tabulated designs is shown in table 3.2-T7. Figures 3.2-44a and 3.2-44b show the frequency responses of the upper and lower band characteristics and also the combined all pass characteristic of the filters.

Due to the relatively large delays of FIR filters at least one design of QMFs using IIR filters has been proposed. This suffers from the effects of group delay distortions and special procedures have to be employed to compensate for this [105].

The special relationships amongst the QM filter coefficients leads to fast implementation algorithms [88, 90]: From equation 3.2-154 the coefficients used for the upper and lower subband filters are identical (except for signs of alternate coefficients). This property can be used to reduce by factor of two the computation load. With regard to figure 3.2-45, partial sums of alternate input samples can be formed, and then added or subtracted together to give the output sample for the lower or upper band respectively.

The above imply a block structure to the algorithm as shown in figure 3.2-46. Here odd numbered samples are weighted and accumulated into one buffer and even samples are accumulated into another. Note that each shift to produce a new block operates at half the original sampling rate. The results from the two blocks are processed through the DFT butterfly to produce the upper and lower band samples. A similar structure can be derived for the receiver as shown in figure 3.2-47. The subband samples are first processed through the DFT butterfly and then channelled to the two separate accumulators to produce samples at the initial sampling rate. An even more efficient approach for implementing QMFs is to use a parallel structure [93]. This is shown to compare favourably with the tree structure described above and should be the choice for real time implementation if the extra amount of execution time saved can be used for another part of the algorithm.

Finally a word should be said about the resulting delay though the filters and the effect of band inversion both for the case of the tree structured QMF.

3.2.3.8a3 Delay

Consider the example shown in figure 3.2-48. Assume that the input signal is sampled at 8Hz, giving samples spaced apart at 1/8 msec each. This is divided at the final stage of the tree into 8 bands of which only the upper branches are shown. Further, assume that each filter has 32 taps. Since the filters are FIR designs they introduce a constant delay of 31/2 samples at the input sampling rate. Since the signals are also decimated the sample spacings at each stage are 1/8 msec at A, 1/4 msec at B, 1/2 msec at C and 1 msec at D. Therefore:

Signal B is delayed by 31/2 samples, spaced at 1/8 msec w.r.t.A

Signal C is delayed by 31/2 samples, spaced at 1/4 msec w.r.t. B

Signal D is delayed by 31/2 samples, spaced at 1/2 msec w.r.t. C

and the total delay between A and D is given by: (31/2)*(1/8 + 1/4 + 1/2) = (31/2)*(7/8) msec = 217/16 msec.

3.2.3.8a4 Frequency Inversion

Consider again, the splitting of the input signal into eight bands as shown in figure 3.2-49, from left to right. At each stage, the lower bands remain unaffected but the spectrum of the upper bands is inverted in frequency. In figure 3.2-49, this is followed through the 3 stages and the final position of the bands due to the successive inversions are shown.

3.2.3.8a5 Transmission rate, SNR and gain over PCM

The transmission rate in SBC over and above that needed for any side information is equal to the sum of the bit rates needed to code the subbands.

In the full band case, assuming a full band width of W, the sampling rate is equal to

$$f_s = 2W \qquad 3.2.156$$

and if R bits/sample are used, the bit rate is equal to

$$B_{full} = f_s R = 2WR \qquad 3.2\text{-}157$$

In the subband case, assuming N equal width bands the sampling frequency is

$$f_{sk} = 2W_k = \frac{2W}{N}$$

each subband sample can be assumed to e coded with R_k bits/sample, therefore the bit rate is

$$B_{sub} = \sum_{k=1}^{N} f_{sk} R_k = \frac{1}{N} \sum_{k=1}^{N} 2WR_k = \frac{1}{N} \sum_{k=1}^{N} R_k . 2W \qquad 3.2\text{-}158a$$

or

$$B_{sub} = 2WR \qquad \text{if} \qquad R = \frac{1}{N} \sum_{k=1}^{N} R_k \qquad 3.2\text{-}158b$$

Therefore a subband coder with equal width contiguous non-overlapping bands of an average bit rate/subband sample of

$$R = \frac{1}{N} \sum_{k=1}^{N} R_k \qquad 3.2\text{-}158c$$

has the same overall bit rate as the equivalent full band coder coded with R bits/sample, with R given by 3.2-158c.

Using the above assumptions of nonoverlapping contiuous equal width bands, the variances of the subband inputs can simply be added to obtain the variance of the full band signal. similarly the variances of subband reconstruction errors can be added together to give the variance of the signal reconstruction error.

Let σ_{qk}^2 be the quantization noise (reconstruction noise) in band k and

$$\sigma_q^2 = \sum_{k=1}^{N} \sigma_{qk}^2 \qquad 3.2\text{-}159$$

be the total reconstruction noise.

Assuming PCM (or (A)DPCM) coding for the bands (although the following formula has a more general application through rate distortion theory):

$$\sigma_{qk}^2 = \epsilon_{*k}^2 2^{-2R_k} \sigma_{xk}^2 \qquad 3.2\text{-}160$$

which is the 6 dB/bit rule.

ϵ^2_{*k} related to quantizer performance (and prediction gain in the case of (A)DPCM) and will be assumed to be the same for each band. σ_{xk}^2 is the signal variance in band K. We aim to minimize the overall noise power by an appropriate bit allocation (Initially the only constraint is that the overall bit rate is constant). Using Lagrange multipliers, the solution to the minimization problem is the solution of

$$\frac{\partial}{\partial R_k}[\sigma_q^2 - \lambda(R - \frac{1}{N}\sum_{k=1}^{N} R_k)] = 0 \qquad 3.2\text{-}161$$

with σ_q^2 given by 3.2-159 and 3.2-160.

It can easily be shown that the solution to the above (constrained) minimization problem is given by

$$R_{k,opt} = R + \frac{1}{2}\log_2 \frac{\sigma_{xk}^2}{[\prod_{l=1}^{N}\sigma_{xl}^2]^{1/N}} \qquad 3.2\text{-}161$$

substituting back into 3.2-160 gives

$$\sigma_{qk}^2 = \epsilon_*^2\, 2^{-2R}\, [\prod_{l=1}^{N}\sigma_{xl}^2]^{1/N} \qquad 3.2\text{-}163$$

Therefore the m.s.e. minimization above results in a flat noise spectrum. Equation 3.2-159 can now be written as

$$\sigma_{qSBC}^2 = N\, \epsilon_*^2 2^{-2R}\, [\prod_{l=1}^{N}\sigma_{xl}^2]^{1/N} \qquad 3.2\text{-}164$$

The noise power from a conventional PCM coder at the same overall bit rate (neglecting any side information rate) is

$$\sigma_{qPCM}^2 = \epsilon_*^2 2^{-2R}\sigma_x^2 \qquad 3.2\text{-}165$$

where σ_x^2 is the total signal variance which can be expressed n terms of the subband variances as

$$\sigma_x^2 = \sum_{k=1}^{N} \sigma_{xk}^2 \qquad 3.2\text{-}166$$

therefore equation 3.2-165 can be written as

$$\sigma_{qPCM}^{2} = \epsilon_{*}^{2} 2^{-2R} \sum_{k=1}^{N} \sigma_{xk}^{2} \qquad 3.2\text{-}167$$

From 3.2-167 and 3.2-164 the SB gain is given by

$$\text{Gain} = \frac{\frac{1}{N} \sum_{k=1}^{N} \sigma_{xk}^{2}}{\left[\prod_{k=1}^{N} \sigma_{xk}^{2} \right]^{1/N}} \qquad 3.2\text{-}168$$

i.e. by the ratio of the arithmetic to the geometric mean of the subband variances and

$$\text{SNR(SBC)}_{dB} = \text{SNR(PCM)}_{dB} + 10\log_{10}(\text{gain}) \text{ dB} \qquad 3.2\text{-}169$$

The minimum value of the subband gain is one when the input spectrum of the full band signal is flat. For a non-flat spectrum gains greater than one can be realized by using the bit allocation formula of 3.2-162. The subband gain is analogous to the prediction gain in (A)DPCM in that they both exploit the non uniform nature of the input spectrum.

The full subband gain will only be realised if each and every R_k in 3.2-162 is positive, since negative R_k is meaningless. This only depends on R. For a sufficiently high value of R all R_k will be positive and the full subband gain can be realized. As R is reduced some of the R_k values will become negative thus reducing the subband gain. This also holds in the use of (A)DPCM coding [78] and is a corrolary of Rate Distortion theory [94]. Various modifications can be made to 3.2-162 to obtain only positive (integer) solutions for R_k [95]. These and also the applications of noise shaping are discussed in Appendix C.

3.2.3.8a6 Speech Coding in Subbands

We will now attempt to follow the historical development of practical realizations of SBC algorithms and provide comparisons with full band coders.

In [86] the explicit method of modulation is described. This enables a subdivision of the signal into bands according to the A.I. Integer band sampling avoids the use of modulators but the splitting of the bands conforms to the A.I. only approximately (within a factor of 2). The subband signals were coded using APCM with the one word memory quantizer of JFC. Two four band designs were developed. In the first design, that for the 16 Kb/s coder, 125 tap overlapping FIR filters were used covering the frequency range from 200-3100 Hz. Three bit coders were used for the two lower bands and two bit coders for the upper bands. This coder was found to give a performance equivalent to a 22 Kb/s ADPCM coder employing a one tap fixed predictor and Jayant's quantizer. For the 9.6 Kb/s coder, nonoverlapping filters were used and small gaps were permitted between the bands. 175-tap FIR filters were used to reduce transition bands and preserve bandwidth. The bit distribution was now 3,2,2,2. This coder was found to be equivalent to 18 kb/s ADM.

Similar results were presented in a later paper [87]. The relationship between SNR per band and subjective preference was established (figure 3.2-50) for four band designs for the low bit rates (7.2-9.6 kb/s) and five band designs for the higher bit rates (16 kb/s). The band divisions are shown on top of figure (3.2-50). Compare this figure with figure 3.2-51 showing the long term spectrum of speech. A relationship between prefered SNR/bad and speech power/band can be seen. The exact relationship is obscured because of the averaging in 3.2-51 and the fixed allocation in 3.2-50.

The measured frequency response of the coders is shown in fig. 3.2-52a,b. Subjective comparisons with other coders revealed that the 16 kb/s SBC was found to be comparable to 26.5 kb/s ADPCM (fixed prediction). The subjective results at the other bit rates were similar to those of ref. [86]

In an attempt to lower the bit rate to 4.8 kb/s, the centre frequency of the two upper bands of a four band scheme was allowed to vary in accordance with the vocal tract resonances F2 and F3. The locations of the resonances were found by zero crossing techniques applied to appropriate subbands of the signal (fig. 3.2-53). Note that this can be considered as the first attempt to a variable bit allocation.

A general comparison of various time domain coders and SBC [96, 97] revealed that a four band SBC gave equivalent performance to an ADPCM coder with an 8th order adaptive predictor (with no noise shaping).

In [91] the QMF designs were introduced and were later tabulated. [92]

A summary of the developments above were presented in [85] in the more general content of frequency domain coding.

In [88] a full description of QMF was given as well as a polyphase structure for fast implementation. Considerations for real time implementations were made in above and in [98].

Adaptive bit allocation based on the instantaneous power in each band was found to give improved results

[99] and simplified algorithms were presented [95]. This necessitates the transmission of side-information which results in an increase of the total bit rate (although the performance is better than a SB coder with fixed bit allocation and at the same total bit rate) and gives rise to error protection implications in the case of transmission over realistic channels. It is interesting to note that in [99] in order to code the suband signals, an ADPCM coder employing an 8th order adaptive (sample to sample) predictor and fixed bit allocation out-performed a design using an adaptive bit allocation and APCM coding in the subjective tests. This was true although the scheme with the adaptive bit allocation gave a higher SNR than the APCM scheme. This was attributed to the band suppression resulting from the adaptive bit allocation which has a small objective effect by definition but significantly affects the perceived speech quality.

Results from real-time implementation reported in [100] include combination of SPC and harmonic scaling, requiring some sort of pitch prediction. Another pitch prediction technique was used in [101]. These techniques attempted to remove both short and long term redundancy from the speech signal. Since the two are quite separate, both lead to improved performance over SBC without pitch prediction. Daumer [102] compared several coders and found pitch predictive ADPCM to be comparable to SBC.

Subband coding is also reported in [103] in an APC coder. A two band split-band scheme with two separate one-tap pitch predictors was found to perform better than the full band 3-tap pitch predictive system.

In [104] a more detailed study of adaptive bit allocation in time and frequency domain is presented for an APC system. This used three subbands, and a fourth order predictor in each subband. The pitch synchronous energy concentration of the residual was exploited by dividing the time domain signals into 4 subintervals and allocating different number of bits to each subinterval. The system provided speech quality subjectively equivalent to 7-bit log-PCM at 16 kb/s and 6 bit log-PCM at 9.6 kb/s. In the above hybrid techniques the performance of SBC can be significantly improved by supplementing it with the time domain operation of adaptive prediction. This removes any within-band redundancies that remain and therefore splits the burden of redundancy removal between the time domain and the frequency domain.

Finally, with the application of Vector Quantization to speech, new coders were developed that combined VQ and SBC In one instance, [106] the subband samples were formed into blocks by synchronously taking one sample from each of the subbands. Note that this technique does not require any side information, and small SNR gains are reported over direct VQ.

In another instance a more familiar design was used [107] of first normalising with the rms of the signal and then again each subband sample by its own rms. The overall rms was log encoded using 5 bits and the block of rms values/band was linearly vector quantized with 8 bits. This vector was used in an adaptive codebook allocation where codebooks of different sizes were allocated to each band for coding the subband signals themselves according to the variances per band. They found that VQ was particularly effective in encoding the side information. A similar system is also described in [108].

3.2.3.8b Adaptive Transform Coding (ATC)

In transform coding systems each block of speech samples is transformed into a set of transform coefficients. These coefficients are then quantized independently and transmitted. An inverse transform is taken at the receiver to obtain the corresponding block of reconstructed speech samples.

Assume that σ_X^2 is the (short term) variance of N successive samples arranged in a vector X. This vector is linearly transformed using a unitary matrix A into a vector Y:

$$Y = A.X \qquad 3.2\text{-}170$$

with

$$A^{-1} = A^T \qquad 3.2\text{-}171$$

The elements of Y are the transform coefficients of the coding scheme. Each of the elements is independently quantized thus leading to a vector $\hat{Y}$. The vector of quantized transform coefficients is transmitted to the receiver and transformed using the inverse matrix A^{-1}.

$$\hat{X} = A^{-1}\hat{Y} \qquad 3.2\text{-}172$$

to obtain the reconstructed output samples forming $\hat{X}$. For unitary matrices the reconstruction error variance is equal to the total quantization error variance:

$$\sigma_q^2 = (X-\hat{X})^T(X-\hat{X}) = (Y-\hat{Y})^T(Y-\hat{Y}) \qquad 3.2\text{-}173$$

by virtue of 3.2-171.

To minimize σ_q^2 an appropriate transform A must be chosen. In addition, the transform coefficients are quantized independently. The variances of the transform coefficients differ in general. R_k bits/sample are allocated for coefficient Y_k of variance σ_{Yk}^2. For an average of R bits/sample overall i.e.

$$R = \frac{1}{N} \sum_{k=1}^{N} R_k \qquad 3.2\text{-}174$$

an optimum bit assignment exists given by

$$R_k = R + \frac{1}{2} \log_2 \frac{\sigma_{yk}^{2}}{\left[\prod_{l=1}^{N} \sigma_{yl}^{2} \right]^{1/N}} \qquad 3.2\text{-}175$$

as in the case of subband coding, although the transform coefficient Y_K replaces subband sample X_K. The transform gain over PCM is given by a similar expression as in the case of subband coding.

The optimum transform is the KLT (Karhunen-Loeve transform). The KLT matrix has the eigen-vectors of the signal's X covariance matrix as its columns. It produces transform coefficients which are uncorrelated.

Since speech is a nonstationary source, a different KLT matrix would have to be calculated for each vector x. This is impractical for a number of reasons and for realistic systems the discrete cosine transform [109] is usually used as the decorrelating matrix A. This provides a performance very close to the KLT transform for speech signals. An important advantage of the DCT as opposed to the KLT transform is that fast algorithms exist for the matrix transformations 3.2-170, 3.2-172 [110, 111, 112]. The transform is defined by

$$Y(k) = \frac{2C(k)}{N} \sum_{j=0}^{N-1} X(j) \cos \left[\frac{(2j+1)k\pi}{2N} \right] \qquad K=0,1,\ldots N-1$$

3.2-176a

and the inverse

$$X(j) = \sum_{k=0}^{N-1} C(k)Y(k)\cos\left[\frac{(2j+1)k\Pi}{2N}\right. \qquad j=0,1...N-1$$

3.2-176b

with

$$C(k) = 1/\sqrt{2} \quad k=0$$

$$C(k) = 1 \qquad k = 1,2....N-1$$

For an adaptive bit assignment, hence a larger prediction gain the (short term) variances $\sigma_{YK}{}^2$ in (3.2-175) need be known at both the transmitter and receiver. These are usually approximated by the squared values of the transform coefficients i.e.

$$\sigma_{yk}{}^2 \simeq Y_k{}^2 \qquad 3.2\text{-}177$$

since in transform coding a very large number of these variances exist due to the large transform sizes (128-256) usually employed a further approximation is used: The values of $Y_K{}^2$ are averaged over a number of neighbouring values and this average is used as the site information representing the $\sigma_{YK}{}^2$ terms of the averaged $Y_K{}^2$. At the receiver the logarithmic averages are interpolated to produce an estimate of $\sigma_{YK}{}^2$. These values, apart form being used to determine the bit allocation, are also used to normalize the Y_K values prior to quantization, a structure resembling AQF. A block diagram of such a coder is shown in figure 3.2-54. The coder provides toll or near toll quality speech around 16 kb/sec [113] but for lower bit rates, clicks, burbling distortion and blockend distortion (i.e. periodic clicks) become audible [114].

A refinement of ATC, using an LPC vocoder model for the site information brings the useful rate for this coder down to about 8 kb/sec. [85] Cepstral models for the side information have also been used [100]. An interesting case arizes when the transform coefficients are grouped together into smaller block-sites and transformed again by smaller transform matrices to produce pseudo-subband time signals as in figure (3.2-55). These pseudo-time signals can then be coded as time signals are, in subband coding. This scheme provides similar performance as SBC without the additional complexity and delay resulting from the filterbank approach of SBC [115].

3.2.3.9 Hybrid Coding: Voice and Residual Excited Vocoders

Waveform coders can provide communication quality speech down to around 9.6 kb/sec. On the other hand vocoders provide synthetic speech around 2.4 kb/sec, with quality saturating as the bit rate is increased beyond around 4.8 b/sec. This leaves a gap in the rage of 4.8-9.6 kb/sec which is filled by hybrid coders sharing some features between waveform coders and vocoders. In perhaps all cases of hybrid coding certain band(s) of the signal are waveform coded whilst the rest of the bands are vocoder driven. They are generally divided into two categories the voice-excited vocoders (VEV) and the residual-excited linear prediction (RELP) vocoders.

3.2.3.9a Voice Excited vocoders

A VEV is illustrated in fig. 3.2-56. A baseband range (typically 0-1000 Hz) is coded as with waveform coders. The excitation signal for the vocoder synthesizer is obtained form the baseband by a process called spectrum-flattening. A spectrum flattener spreads out the spectrum of the baseband signal by nonlinear distortion to cover the frequency band to be synthesized; then the frequency components generated by the nonlinear distortion are equalized to form a flat-spectrum excitation signal. The equalization can be achieved "instantaneously" by means of a bank of contiguous band-pass filters to which the signal is applied followed by hard limiters ("infinite clippers"). Alternatively, fast automatic gain controls can be used. A spectrum flattener using clippers is shown in figure 3.2-57. Further, the output of the clippers can be made proportional to the corresponding spectral amplitude of the original speech signal. The second column of band-pass filters removes nonlinear distortion components introduced by the hard limiters. The excitation signal in VEV has inherently the correct periodicity; for an aperiodic input the output of the spectrum-flattener is also aperiodic and, for a periodic input, the output will periodic with the same periodicity.

3.2.3.9b <u>Residual Excited Linear Prediction (RELP) coders</u>

The operations required for a flat excitation and performed through the spectral flatteners can be applied to the LPC residual instead of the speech signal. Such an approach is shown in figure 32-58. Note that this is an improvement to the LPC vocoder whereas the former system described offered an improvement to the channel vocoder. The baseband can be coded through any of the waveform coding approaches already mentioned, such as APC [117], SBC [118], ATC [119].

The high frequency regeneration HFR can be performed through Rectification [120], spectral flattening as in the VEV described above, or through more complicated clipping characteristics [121], pitch-controlled spectral shifting [119] or by spectral duplication [120], [116].

The latter method above takes advantage of the fact that the baseband residual is already flat. In this regeneration method, the aim is to simply duplicate the baseband spectrum at higher frequencies in some fashion. This can be done in the frequency domain as in [119]. In the time domain, similar results can be achieved. Assume that the signal bandwidth is L times the base-band bandwidth i.e. W/B = L where L is an integer. Two different methods can be distinguished, spectral folding and spectral translation, figures 3.2-59 b and c respectively. The baseband is shown in 3.2-59 and L = 3. In figure 3.2-59 the spectrum in the second band (between B and 2B) is the mirror image (folded version) of the baseband, while the spectrum in the third band is a folded version of the spectrum in the second band, hence identical to the spectrum of the first band (baseband). In figure 3.2-59c the second and third bands have spectra identical to the baseband. The spectra are obtained by translating (and copying) the baseband.

The receivers for both cases are shown in figure 3.2-60 (for folding) and 3.2-61 (translation). In figure 3.2-60 the introduction of L-1 zeros after each sample creates the necessary images in a way identical to integer band filter banks in subband coding. In figure 3.2-61 the multiplication of every other sample by -1 inverts the spectrum. The filter H(Z) then passes the intervening bands (frequency inverted). The quality of the regenerated speech improves if the short term dc of the baseband signal is subtracted prior to regeneration. This can be added to the signal after the regeneration process [120, 116]. Also perturbed spectral folding improves the quality [116]. This is achieved by perturbing the non zero samples of the upsampled residual that are below a certain threshold (in order not to disturb the pitch structure).

Note that the above frequency duplication only assures that the spacing of harmonics is maintained in the higher bands but not their positions, thus creating in a sense an inharmonic signal. The perceptual impact of this is known to be small as was shown in the chapter on Hearing (under pitch perception). A way to avoid shifting the pitch harmonics is to apply pitch prediction [122] or shift the harmonics back to their original place through additional side information [119].

The coders under the multipulse-LPC structure [79] and Regular Pulse - LPC structure [81] can be thought to represent a generalized approach to baseband coding. As in RELP with spectral folding (figure 3.2-60) an upsampled signal is used to excite an LPC filter, for the case of Regular Pulse - LPC, and a perturbed upsampled signal [116] is used in multipulse LPC. The equivalent "baseband" signal though is not a lowpass version of the excitation as in RELP coders but a more complicated signal [81].

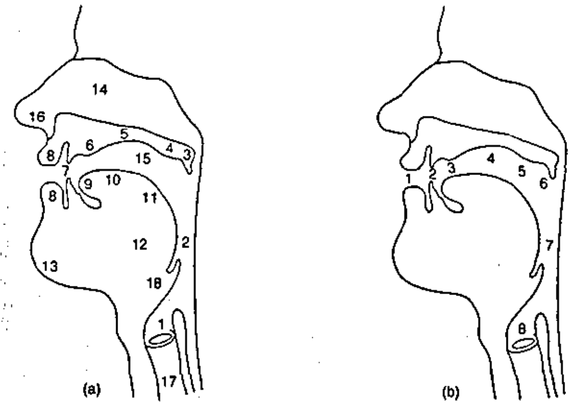
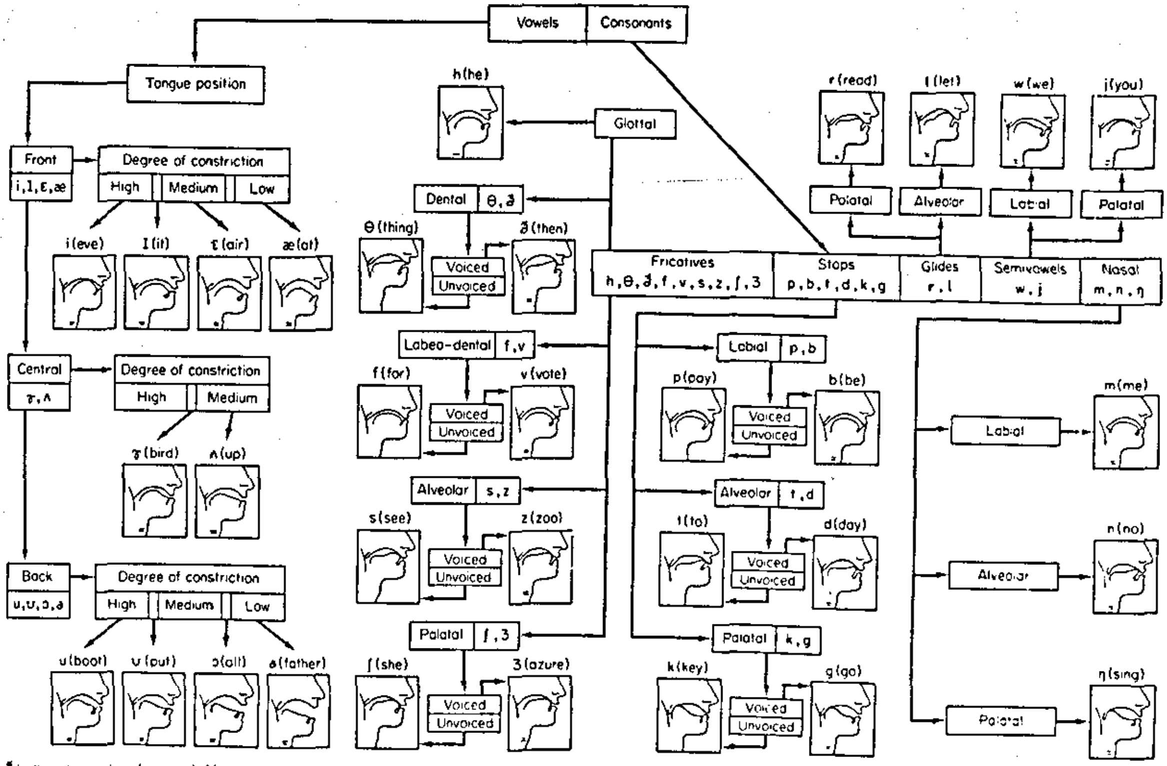


Figure 3.1-1 [1] A cross-sectional view of the vocal tract. (a) Speech articulators: (1) vocal folds, (2) pharynx, (3) velum, (4) soft palate, (5) hard palate, (6) alveolar ridge, (7) teeth, (8) lips, (9) tongue tip, (10) blade, (11) dorsum, (12) root, (13) mandible (jaw), (14) nasal cavity, (15) oral cavity, (16) nostrils, (17) trachea, (18) epioglottis. (b) Places of articulation (1) labial, (2) dental, (3) alveolar, (4) palatal, 5) velar, (6) uvular, (7) pharyngeal, (8) glottal.



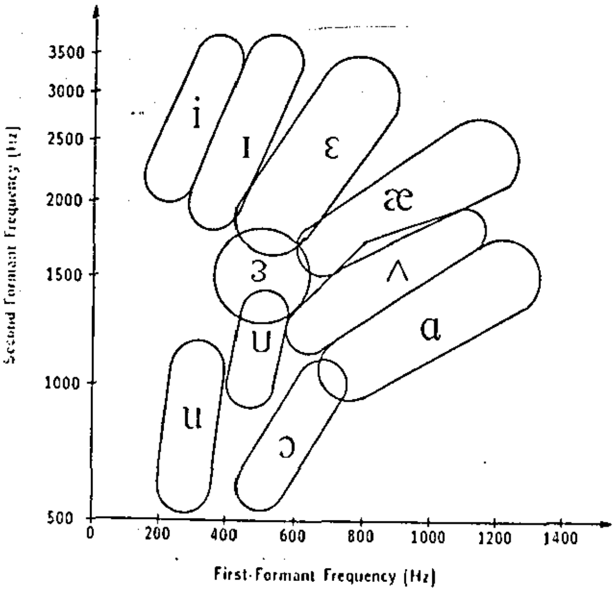
* In the above classification /e/ (hate), /o/ (obey) are considered as diphthongs and /tʃ/ (chew) and /dʒ/ (jar) are considered as stop-fricative combinations

Classification of phonemes according to their manner and place of production.

Figure 3.1-2 [Ref. d, Ch. 2]

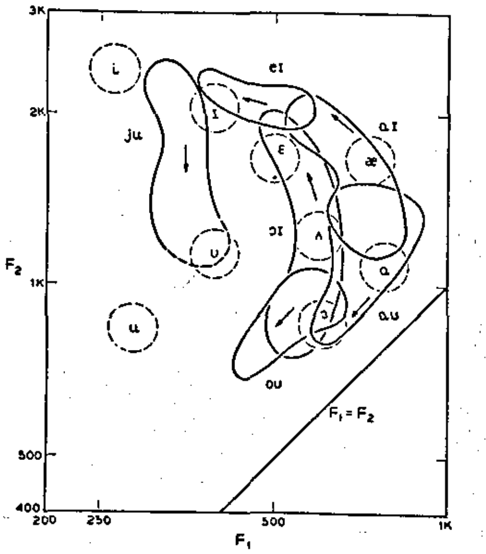
Phoneme	Manner of Articulation	Place of Articulation	Voiced	Example Word
i	vowel	high front tense	yes	beat
I	vowel	high front lax	yes	bit
e	vowel	mid front tense	yes	bait
ϵ	vowel	mid front lax	yes	bet
æ	vowel	low front tense	yes	bat
α	vowel	low back tense	yes	cot
⊃	vowel	mid back lax rounded	yes	caught
o	vowel	mid back tense rounded	yes	coat
U	vowel	high back lax rounded	yes	book
u	vowel	high back tense rounded	yes	boot
ʌ	vowel	mid back lax	yes	but
ɜ	vowel	mid tense (retroflex)	yes	curt
ə	vowel	mid lax (schwa)	yes	*a*bout
αj (αI)	diphthong	low back → high front	yes	bite
⊃j (⊃I)	diphthong	mid back → high front	yes	boy
αw (αU)	diphthong	low back → high back	yes	bout
j	glide	front unrounded	yes	you
w	glide	back rounded	yes	wow
l	liquid	alveolar	yes	lull
r	liquid	retroflex	yes	roar
m	nasal	labial	yes	maim
n	nasal	alveolar	yes	none
η	nasal	velar	yes	bang
f	fricative	labiodental	no	fluff
v	fricative	labiodental	yes	valve
θ	fricative	dental	no	thin
δ	fricative	dental	yes	then
s	fricative	alveolar strident	no	sass
z	fricative	alveolar strident	yes	zoos
ʃ	fricative	palatal strident	no	shoe
	fricative	palatal strident	yes	measure
h	fricative	glottal	no	how
p	stop	labial	no	pop
b	stop	labial	yes	bib
t	stop	alveolar	no	tot
d	stop	alveolar	yes	did
k	stop	velar	no	kick
g	stop	velar	yes	gig
č	affricate	alveopalatal	no	church
ž	affricate	alveopalatal	yes	judge

Table 3.1-T1 [1] English phonemes and corresponding features.



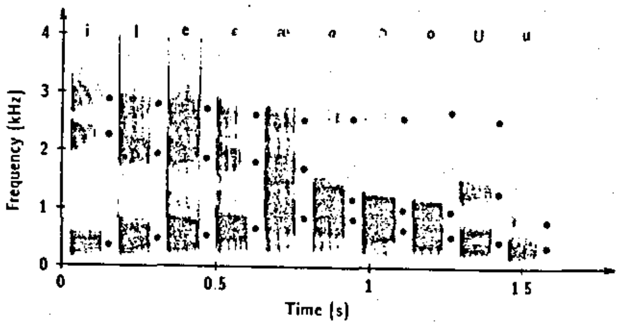
Plot of F1 vs. F2 for vowels spoken by 60 speakers.

Figure 3.1-3 [3]



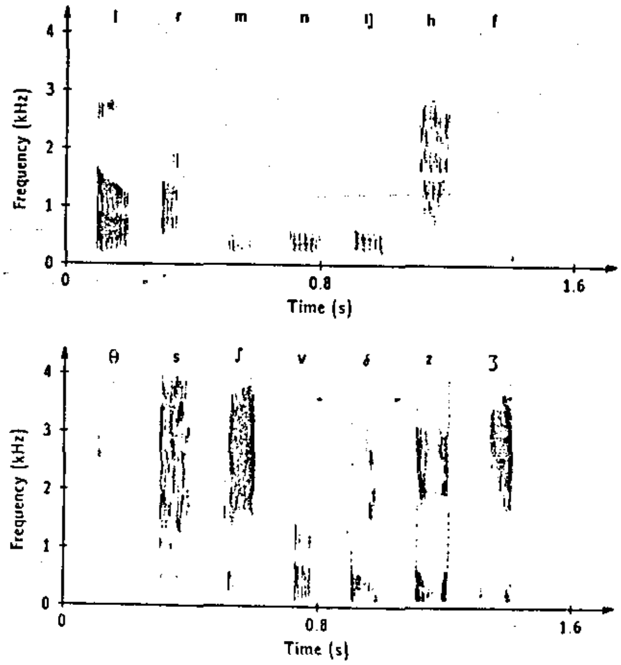
Time variations of the first two formants for diphthongs.

Figure 3.1–4 [4]



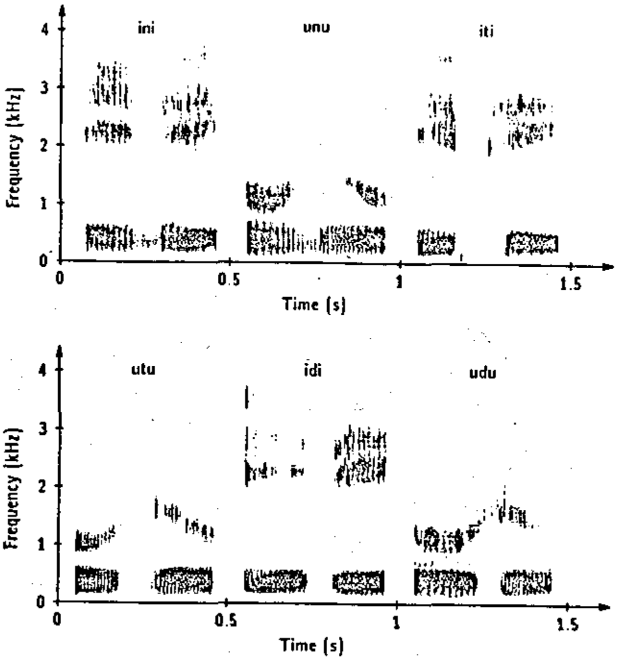
Spectrogram of short sections of English vowels from a male speaker. Formants for each vowel are noted by dots.

Figure 3.1–5 [1]



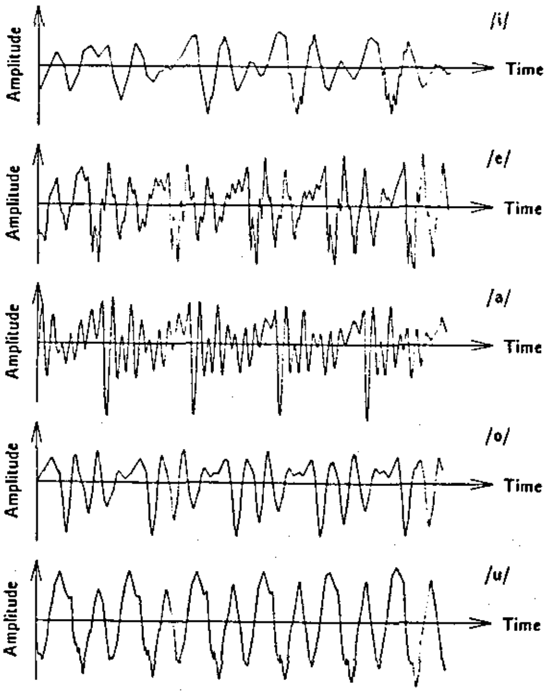
Spectrograms of 14 English steady-state consonants /l,r,m,n,ŋ,h,f, θ,s,ʃ,v,ð,z,ʒ/.

Figure 3.1–6 [1]



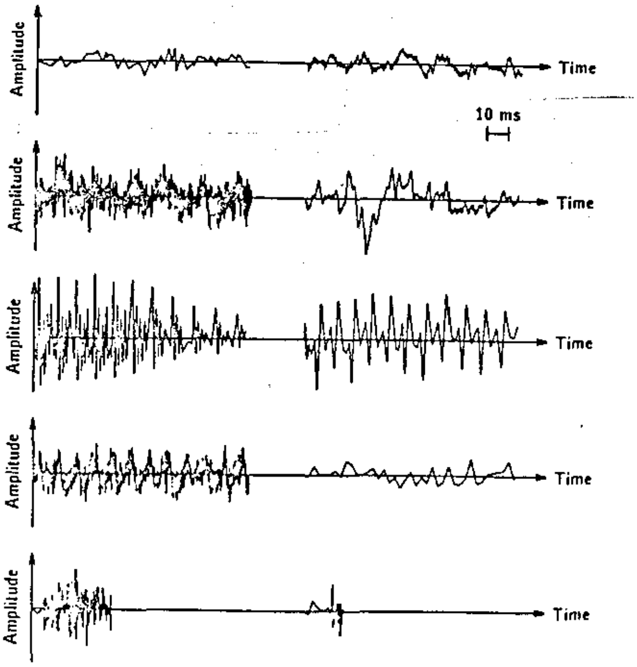
Spectrograms of English stops and nasals in vocalic context: /ini,unu,iti,utu,idi,udu/.

Figure 3.1-7 [1]



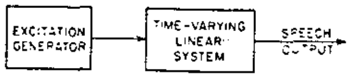
Typical acoustic waveforms for five English vowels. Each plot shows 40 ms of a different vowel, which comprises about 5-6 pitch periods for this speaker. Note the quasi-periodic nature of such voiced speech as well as the varying spectral content for different vowels.

Figure 3.1-8 [1]



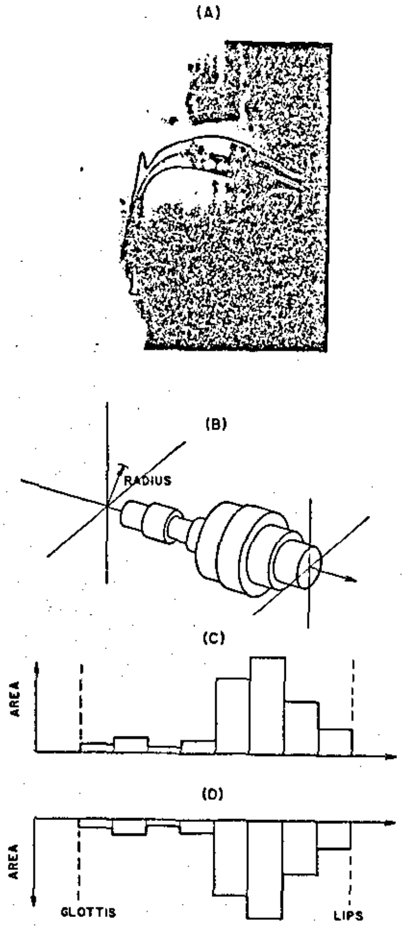
Typical acoustic waveforms for ten English consonants /l,n,ʒ,v,k,g,f,s,ʃ,z/.

Figure 3.1-9 [1]



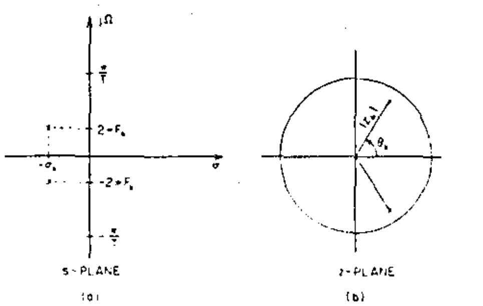
Source-system model of speech production.

Figure 3.1-10 [2]



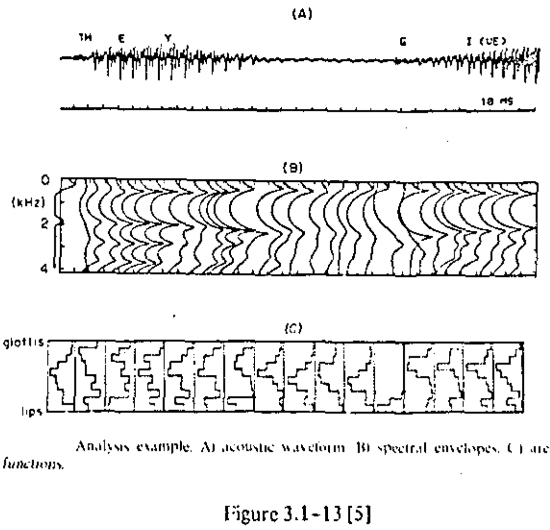
Vocal tract representations. A) a mid-sagittal X-ray section. B) series of uniform cylindrical sections. C) discrete area function. D) area function with reversed axis.

Figure 3.1-11 [5]



(a) s-plane; and (b) z-plane representations of a vocal tract resonance.

Figure 3.1-12 [2]



Analysis example. A) acoustic waveform. B) spectral envelopes. C) area functions.

Figure 3.1-13 [5]

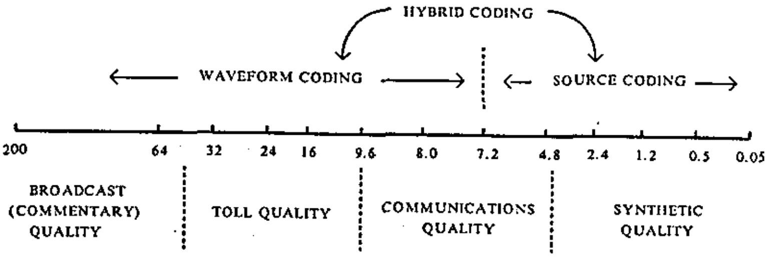
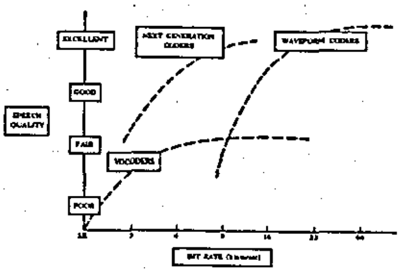
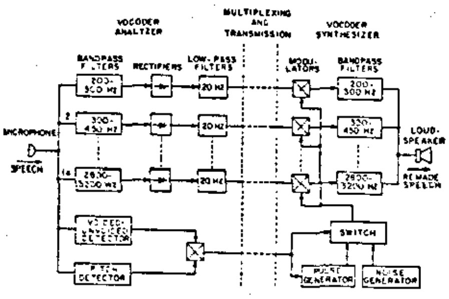


Figure 3.2-1 Spectrum of speech coding transmission rates (nonlinear scale) in Kbits/sec and associated quality. Adapted form [9].



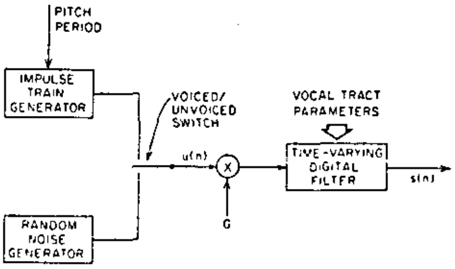
Speech quality versus bit rate for different types of coders.

Figure 3.2-2 [10]



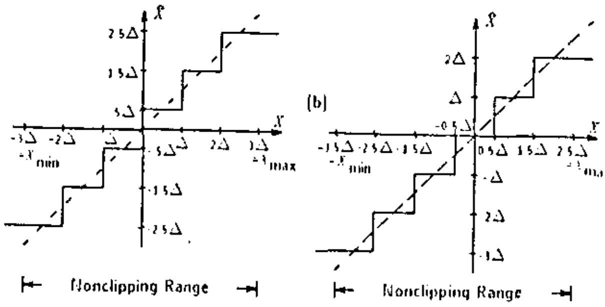
Block diagram of a channel vocoder.

Figure 3.2-3 [9]



Block diagram of simplified model for speech production.

Figure 3.2-4 [2]



Two common uniform quantizer characteristics: (a) mid-riser. (b) mid-tread. The dashed line indicates the ideal curve.

Figure 3.2-5 [1]

Table 3.2-T1 [21] Optimum step size and max{SNR} for uniform symmetric quantizers with different input pdf's (U: Uniform G: Gaussian L: Laplacian Γ: Gamma)

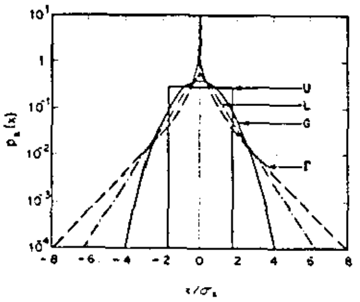
R (bits/sample)	Δ_{opt}/σ_x				max {SNR} (dB)			
	pdf				pdf			
	U	G	L	Γ	U	G	L	Γ
1	1.7320	1.5956	1.4142	1.1547	6.02	4.40	3.01	1.76
2	0.8660	0.9957	1.0874	1.0660	12.04	9.25	7.07	4.95
3	0.4330	0.5860	0.7309	0.7957	18.06	14.27	11.44	8.78
4	0.2165	0.3352	0.4610	0.5400	24.08	19.38	15.96	13.00
5	0.1083	0.1881	0.2800	0.3459	30.10	24.57	20.60	17.49
6	0.0541	0.1041	0.1657	0.2130	36.12	29.83	25.36	22.16
7	0.0271	0.0569	0.0961	0.1273	42.14	35.13	30.23	26.99
8	0.0135	0.0308	0.0549	0.0743	48.17	40.34	35.14	31.89

Table 3.2-T2 [21] *Four model pdf's with a mean value of zero.*

Name	Notation	$p_x(x)$
Uniform or Rectangular	U $(\sigma_x^2 = \Delta^2/12)$	$\frac{1}{\Delta}$; $x \in (-\frac{\Delta}{2}, \frac{\Delta}{2})$ 0: *otherwise*
Gaussian or Normal	G $N(0,\sigma_x^2)$	$\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\{-x^2/2\sigma_x^2\}$
Laplacian or Two-sided Exponential	L	$\frac{1}{\sqrt{2}\sigma_x}\exp[-\sqrt{2}\lvert x\rvert/\sigma_x]$
Gamma	Γ	$\frac{\sqrt[4]{3}}{\sqrt{8\pi\sigma_x\lvert x\rvert}}\exp\{-\sqrt{3}\lvert x\rvert/2\sigma_x\}$

Table 3.2-T3 [21] Optimum decision values x_j and reconstruction values y_j for pdf-optimized nonuniform quantizers (U: Uniform G: Gaussian L: Laplace Γ: Gamma). *Note that in this table, the quantizer characteristics are symmetrical about zero;* $j = 1$ corresponds to the first non-negative value of x or y, and $j > 1$ to succeeding positive values.

R		1		2		3		4	
pdf	j	$x_{j,opt}$	$y_{j,opt}$	$x_{j,opt}$	$y_{j,opt}$	$x_{j,opt}$	$y_{j,opt}$	$x_{j,opt}$	$y_{j,opt}$
U	1	0.000	0.866	0.000	0.433	0.000	0.217	0.000	0.109
	2			0.866	1.299	0.433	0.650	0.217	0.326
	3					0.866	1.083	0.433	0.542
	4					1.299	1.516	0.650	0.759
	5							0.866	0.975
	6							1.083	1.192
	7							1.299	1.408
	8							1.516	1.624
G	1	0.000	0.798	0.000	0.453	0.000	0.245	0.000	0.128
	2			0.982	1.510	0.501	0.756	0.258	0.388
	3					1.050	1.344	0.522	0.657
	4					1.748	2.152	0.800	0.942
	5							1.099	1.256
	6							1.437	1.618
	7							1.844	2.069
	8							2.401	2.733
L	1	0.000	0.707	0.000	0.420	0.000	0.233	0.000	0.124
	2			1.127	1.834	0.533	0.833	0.264	0.405
	3					1.253	1.673	0.567	0.729
	4					2.380	3.087	0.920	1.111
	5							1.345	1.578
	6							1.878	2.178
	7							2.597	3.017
	8							3.725	4.432
Γ	1	0.000	0.577	0.000	0.313	0.000	0.155	0.000	0.073
	2			1.268	2.223	0.527	0.899	0.230	0.387
	3					1.478	2.057	0.591	0.795
	4					3.089	4.121	0.051	1.307
	5							1.633	1.959
	6							1.390	2.822
	7							3.422	4.061
	8							5.128	6.195



Four model pdf's with a mean value of zero

Figure 3.2-6 [21]

Performance of pdf-optimized quantizers. Values of max[*SNR*] (in dB). (U: Uniform G: Gaussian L: Laplace Γ: Gamma)

R (bits per sample)	pdf			
	U	G	L	Γ
1	6.02	4.40	3.01	1.76
2	12.04	9.30	7.54	6.35
3	18.06	14.62	12.64	11.52
4	24.08	20.22	18.13	17.07
5	30.10	26.01	23.87	22.85
6	36.12	31.89	29.74	28.73
7	42.14	37.81	35.69	34.67

Table 3.2–T4 [21]

Figure 3.2–7 [21] The use of a compression function $c(x)$ to realize a nonuniform quantizer characteristic.

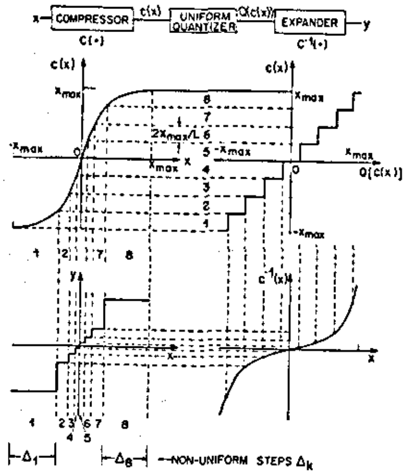


Figure 3.2–8 [21] Dependence of *SNR* on σ_x^2 in 8-bit log- and uniform quantization of a bounded Laplacian input. The dashed line refers to a pdf-optimized quantizer.

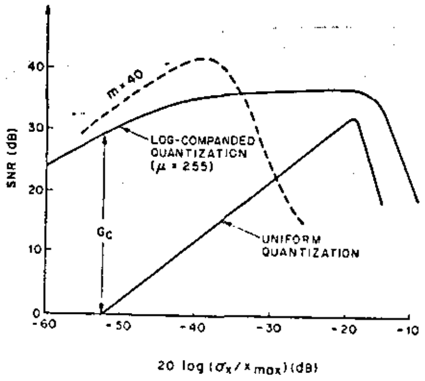
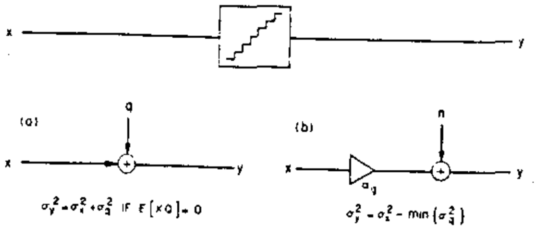


Figure 3.2–9 [21] Quantization noise models: (a) purely additive noise model; and (b) model with non-unity gain and additive noise.



x
y
(a)
q
x
+
y
$\sigma_y^2 = \sigma_x^2 + \sigma_q^2$ IF $E[xq] = 0$
(b)
n
x
a_g
+
y
$\sigma_y^2 = \sigma_x^2 - \min\{\sigma_q^2\}$

Figure 3.2-10 [21] Adaptive quantization with (a) forward estimation of input level (AQF), and (b) backward estimation of input level (AQB).

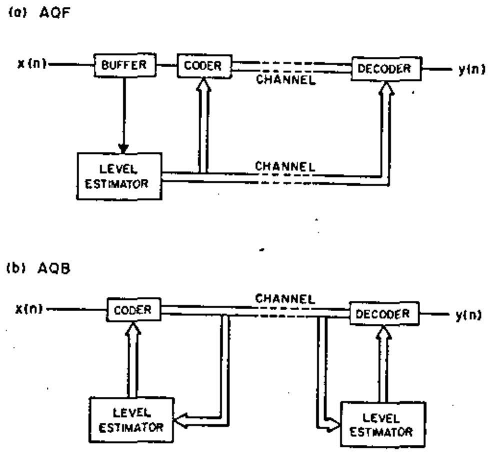
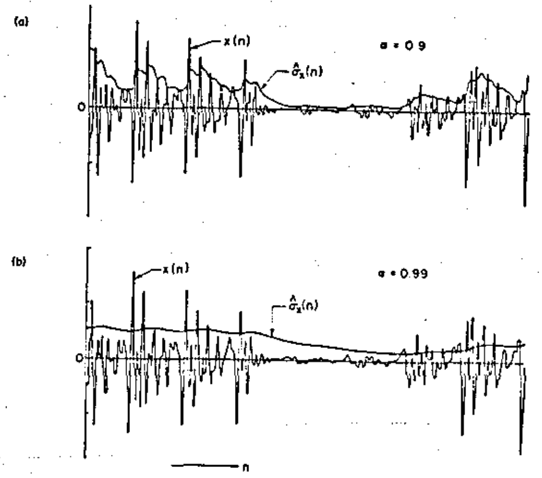


Figure 3.2-11 [24] Speech variance estimates in (a) instantaneous and (b) syllabic adaptation. Values of $\hat{\sigma}_x(n)$ have been magnified for clarity, but the scaling factors are the same in (a) and (b) [Barnwell et al., 1974. Reprinted with permission].



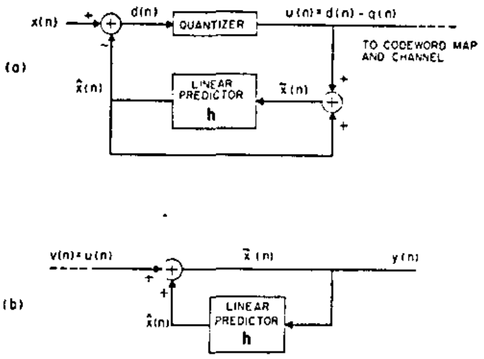


Figure 3.2-12 [21] Block diagram of DPCM: (a) coder; and (b) decoder.

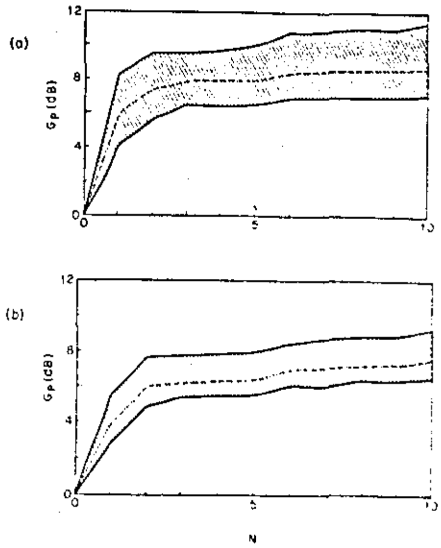


Figure 3.2-13 [21] Maximum prediction gain versus predictor order N for (a) lowpass-filtered speech; and (b) bandpass-filtered speech

Figure 3.2–14 [26] Comparison of log-PCM and DPCM-AQB speech coders. Log spectra of (a) speech input and of reconstruction errors in (b) DPCM-AQB (R = 4 bits/sample) and (c) log-PCM (R = 7 bits/sample)

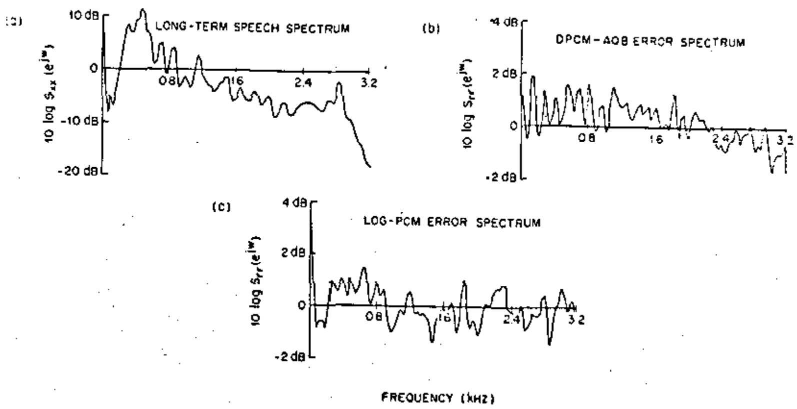


Figure 3.2–15 [21] Block diagram of DPCM with (a) forward-adaptive prediction (APF); and (b) backward-adaptive prediction (APB).

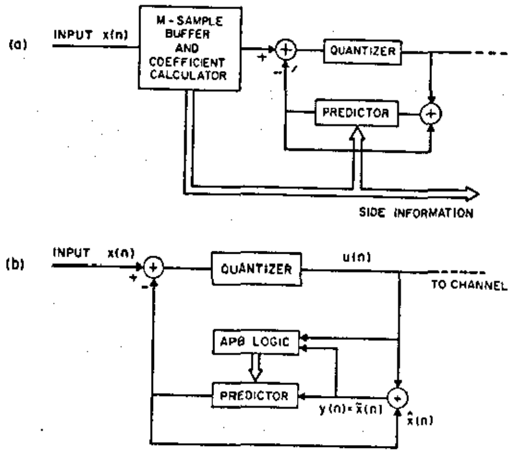


Figure 3.2–16 [21] Maximum prediction gain versus order N of adaptive predictor for (a) lowpass-filtered speech; and (b) bandpass-filtered speech

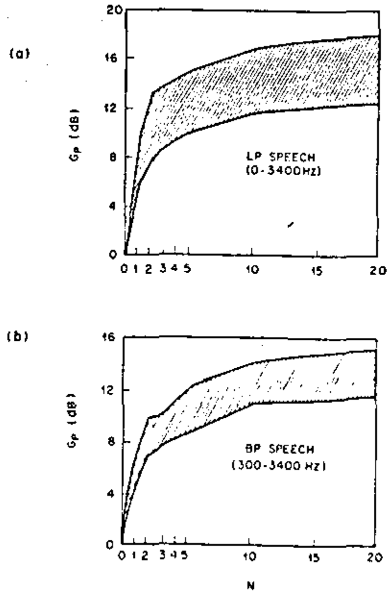


Figure 3.2-17 [21] Time dependencieś of (a) input speech level and of prediction gain G_p in (b) nonadaptive prediction and (c) adaptive prediction. All three waveforms are 1440 ms long and are sampled once every 16 ms. The points M, N and N refer to nasal sounds (in the word "München"), which are characterized by very high values of G_p, in the order of 20 dB [Noll, 1973].

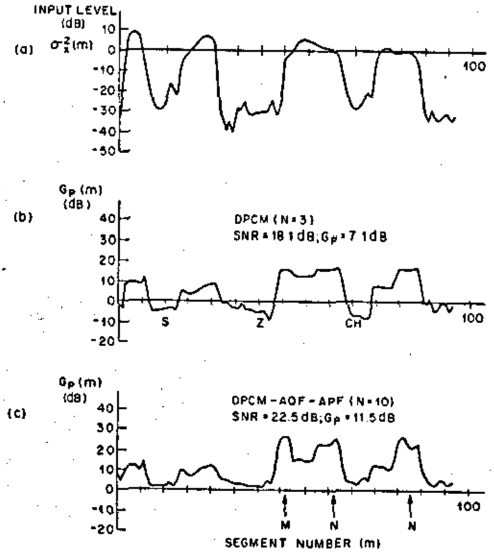


Table 3.2-T5 [26] Comparison of Objective and Subjective Performance of ADPCM and Log-PCM.

Objective Rating (SNR)	*Subjective Rating (Preference)*
7-bit PCM	7-bit PCM (High Preference)
6-bit PCM	4-bit ADPCM
4-bit ADPCM	6-bit PCM
5-bit PCM	3-bit ADPCM
3-bit ADPCM	5-bit PCM
4-bit PCM	4-bit PCM (Low Preference)

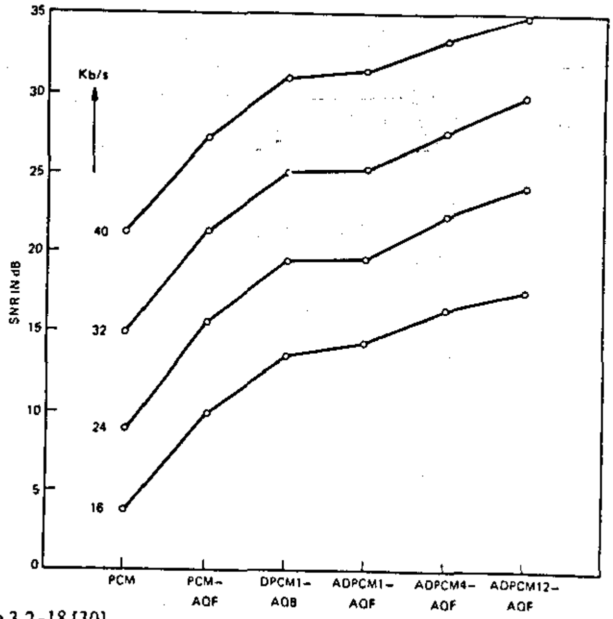
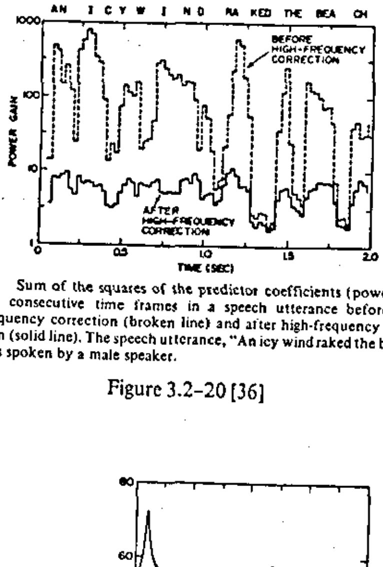
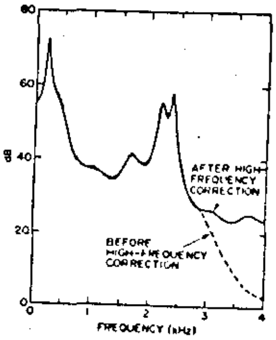


Figure 3.2-18 [30] Signal-to-noise ratio values for quantization with two bits per sample (16 kb/s) up to five bits per sample (40 kb/s). Code: AQF - Adaptive quantizer - feed forward; AQB - Adaptive quantizer - feed backward; ADPCM, - ADPCM system with r^{th} order predictor.



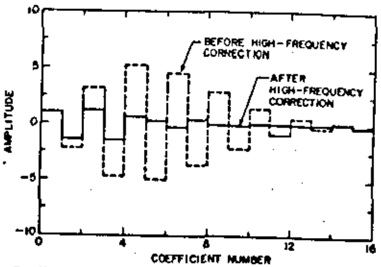
Sum of the squares of the predictor coefficients (power gain) for consecutive time frames in a speech utterance before high-frequency correction (broken line) and after high-frequency correction (solid line). The speech utterance, "An icy wind raked the beach," was spoken by a male speaker.

Figure 3.2-20 [36]



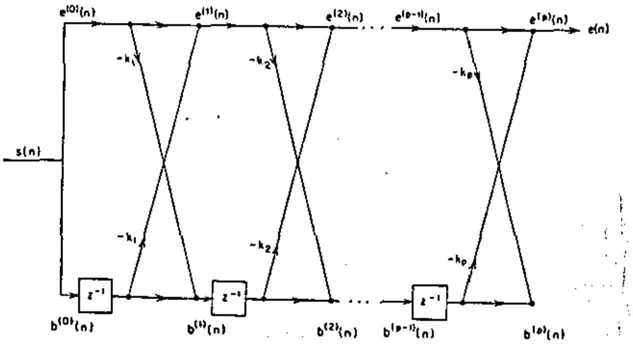
Spectral envelopes of speech based on LPC analysis before high-frequency correction (solid curve) and after high-frequency correction (dotted curve).

Figure 3.2-19 [36]



Predictor coefficients from LPC analysis of speech before high-frequency correction (broken lines) and after high-frequency correction (solid lines).

Figure 3.2-21 [36]



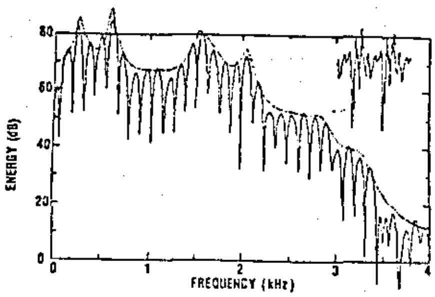
Block diagram of a realizable implementation of the lattice method.

Figure 3.2-22 [2]

Computational Considerations in the LPC Solutions

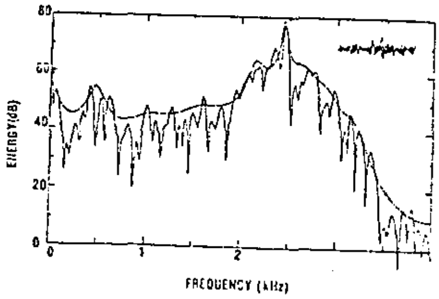
	Covariance Method	*Autocorrelation Method*	*Lattice Method*
	(Cholesky Decomposition)	(Durbin Method)	(Burg Method)
Storage			
Data	N_1	N_2	$3N_3$
Matrix	proportional to $p^2/2$	proportional to p	—
Window	0	N_2	—
Computation (Multiplications)			
Windowing	0	N_2	—
Correlation	proportional to N_1p	proportional to N_2p	—
Matrix Solution	proportional to p^3	proportional to p^2	$5N_3p$

Table 3.2-T6 [2]



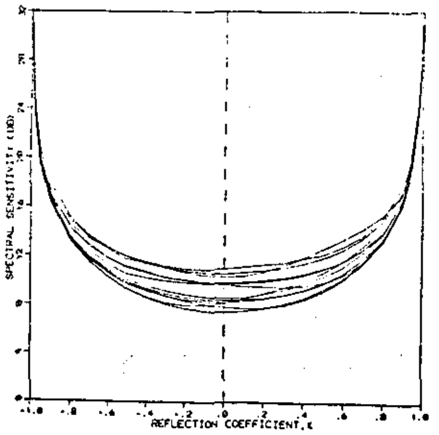
A short segment of a voiced speech signal (upper right inset) is Fourier transformed to yield this logarithmic power spectrum. Note the regular harmonic structure of the spectrum at multiples of the fundamental frequency (ca. 120 Hz). The dotted line is the spectral envelope computed from the frequency response of the prediction filter. - Short-time spectra, such as those shown in Figures 2 and 3, have a high frequency resolution in spite of the shortness of the time window. Following a suggestion by B.S. Atal, this feat is accomplisned by performing the Fourier transform on the spectrally flattened speech signal, thereby minimizing the frequency uncertainties ("splatter") resulting from short time windows.

Figure 3.2-23 [49]



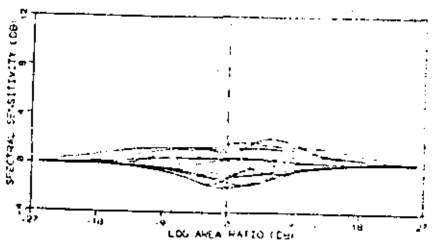
A short segment of an unvoiced speech signal (upper right inset) and its Fourier transform, obtained by the method described in the caption for Figure 2. Note the absence of harmonic frequencies and the broad spectral peak, characteristic of unvoiced sounds. Noise-free coding of such hiss-like sounds is not critical, because "noise plus noise equals noise" and the ear is somewhat more tolerant to spectral distortions of fricative speech sounds.

Figure 3.2-24 [49]



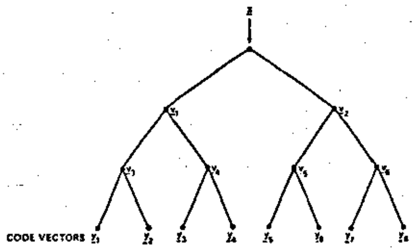
Typical spectral sensitivity curves for the reflection coefficients of a 12-pole analysis of a 20 ms speech frame.

Figure 3.2-25 [50]



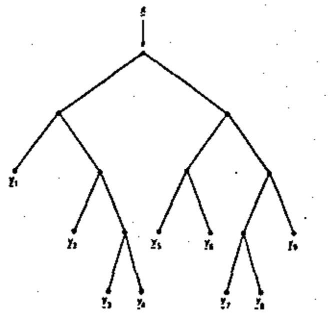
Spectral sensitivity curves using the log area ratios

Figure 3.2-26 [50]



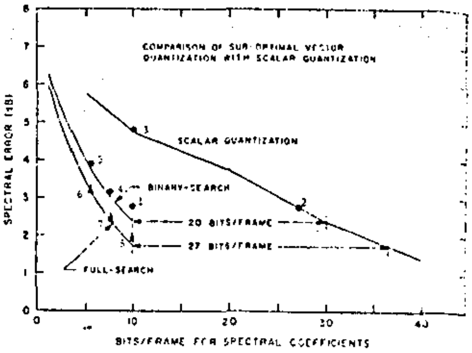
Uniform tree for a binary-search vector quantizer. The vectors v_i are intermediate code vectors that are compared with the input vector x. The code vectors are the vectors y_i. The codebook size L is restricted to be a power of 2 in a uniform binary search.

Figure 3.2-27 [52]



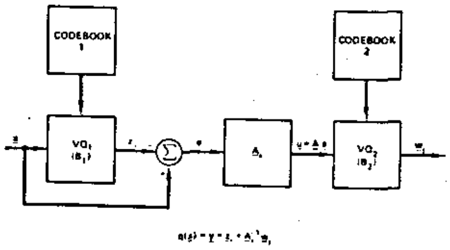
Nonuniform tree for a binary-search vector quantizer. The codebook size in this case can be any integer.

Figure 3.2-28 [52]



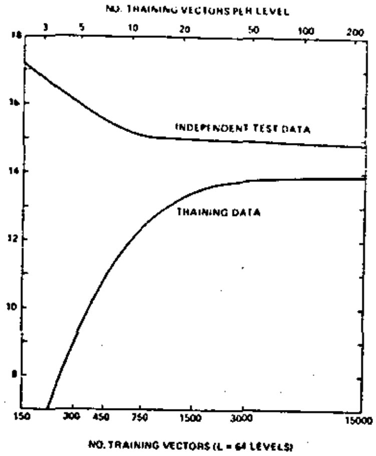
Comparison of suboptimal vector quantization with scalar quantization.

Figure 3.2-29 [54]



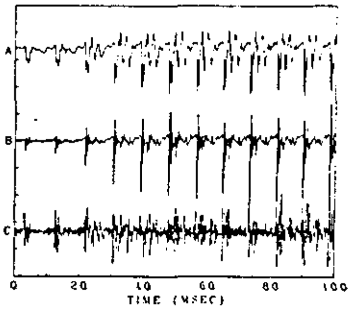
A two-stage cascaded vector quantizer. In the first stage, a B_1-bit vector quantizer quantizes x into a particular vector z_i. The "residual" vector $e = x - z_i$ is then quantized using a B_2-bit quantizer. (The rotation A is used to "realign" all the residual vectors to reduce the distortion.) The quantized value of x is then given as the sum of the two code vectors, as shown in the figure.

Figure 3.2-30 [52]



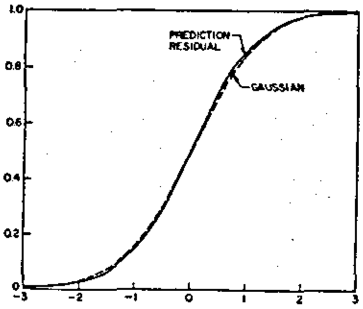
The mse in quantizing LARs for a 6-bit codebook (L = 64 levels) as a function of the number of training data vectors, when tested on the training data and on an independent data set. The smaller the gap between the two curves, the more robust is the codebook.

Figure 3.2–31 [52]



(A) Speech waveform. (B) Difference signal after prediction based on spectral envelope (amplified 10 dB relative to the speech waveform). (C) Difference signal after prediction based on pitch periodicity (amplified 20 dB relative to the speech waveform).

Figure 3.2–32 [36]



First-order cumulative amplitude distribution function for the prediction residual samples (solid curve). The corresponding Gaussian distribution function with the same mean and variance is shown by the dashed curve.

Figure 3.2–33 [36]

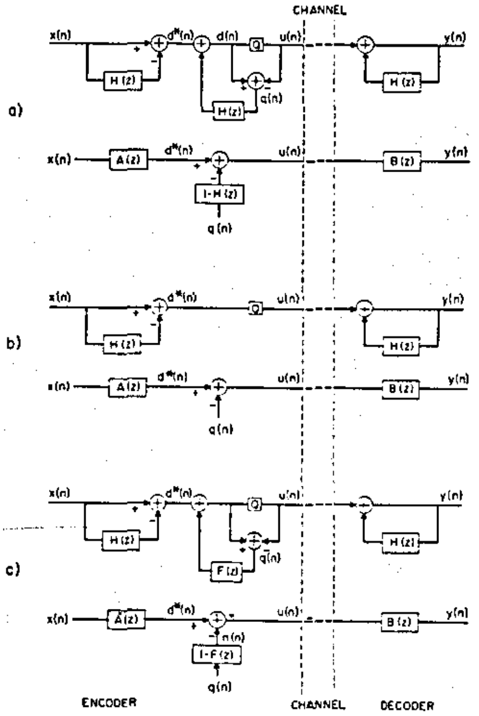
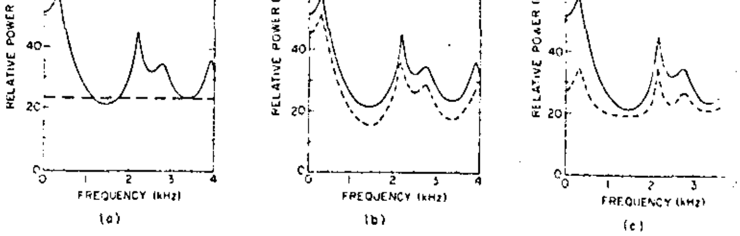
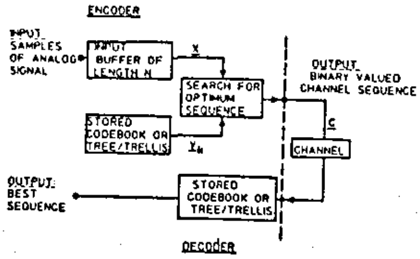


Figure 3.2–34 [21] (a) An alternative description of the DPCM circuit (b) The special case (D*PCM) obtained by eliminating noise feedback. (c) The general noise feedback coder (NFC) obtained by replacing the feedback filter $H(z)$ in (a) by a more general feedback filter $F(z)$.

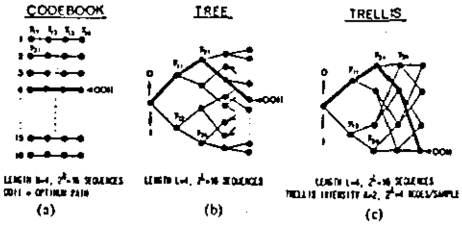
Figure 3.2–35 [35] Comparisons of coder input spectrum (solid curves) and coding noise spectrum (dashed curves) for the three coders of Figure 7.1: (a) DPCM; (b) D*PCM; and (c) NFC. The input is segment of voiced speech and $H_{opt}(z)$ is the transfer function of the mmse predictor for this input




ENCODER
INPUT SAMPLES OF ANALOG SIGNAL
INPUT BUFFER OF LENGTH N
X
SEARCH FOR OPTIMUM SEQUENCE
STORED CODEBOOK OR TREE/TRELLIS
Y_k
OUTPUT BINARY VALUED CHANNEL SEQUENCE
C
CHANNEL
OUTPUT BEST SEQUENCE
STORED CODEBOOK OR TREE/TRELLIS
DECODER

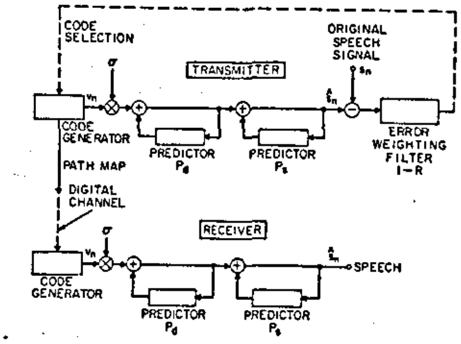
Structure of a multipath search coding (MSC) scheme.

Figure 3.2-36 [70]



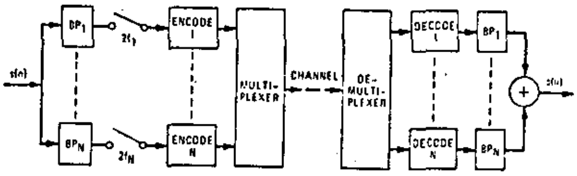
MSC coding classes.

Figure 3.2-37 [70]



Block diagram of a tree coder for speech signals using adaptive source and error-weighting filters.

Figure 3.2-38 [77]



Implementation of a sub-coder based on integer-band sampling.

Figure 3.2-39 [87]

Amplitude responses in filter-banks consisting of four individual bandpass characteristics of (a) equal width and (b) unequal width.

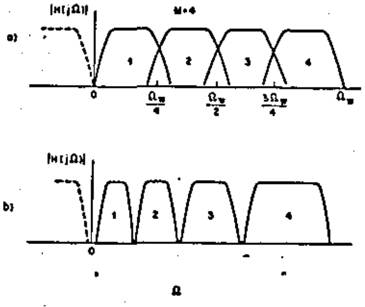
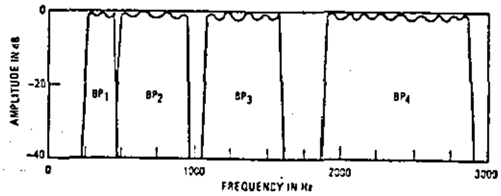
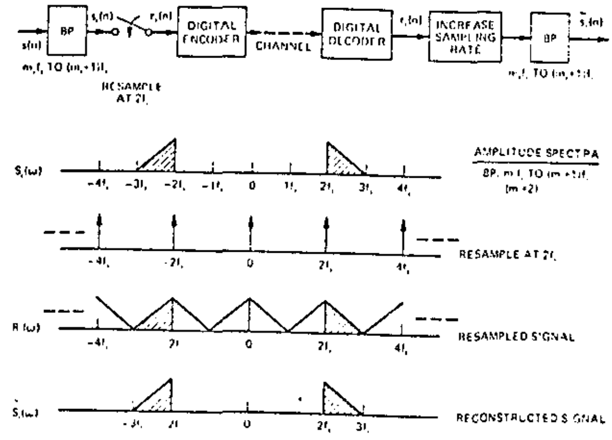


Figure 3.2-40 [21]



Frequency-domain illustration of the sub-band partitioning of the speech band.

Figure 3.2-41 [87]



Integer-band sampling technique and a frequency-domain interpretation.

Figure 3.2-42 [87]

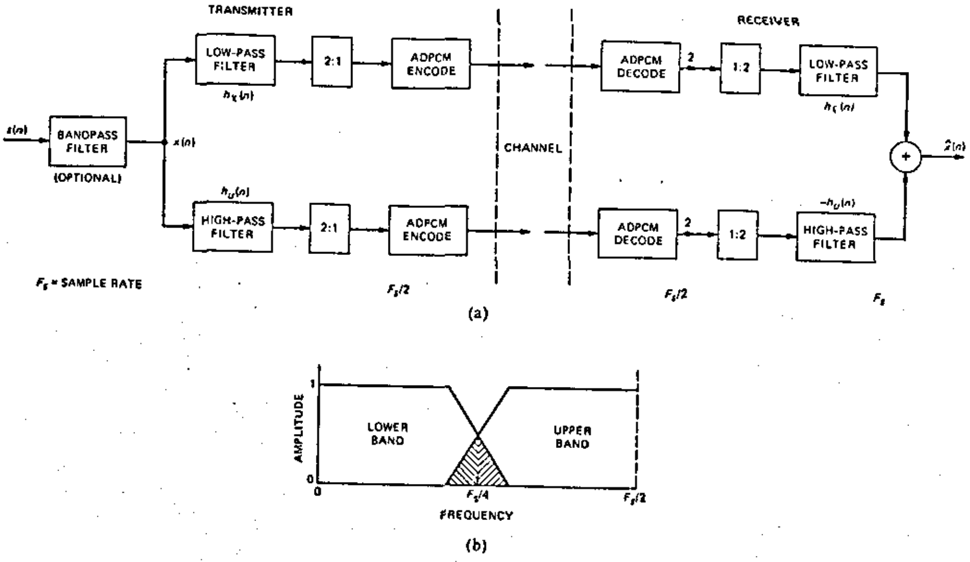
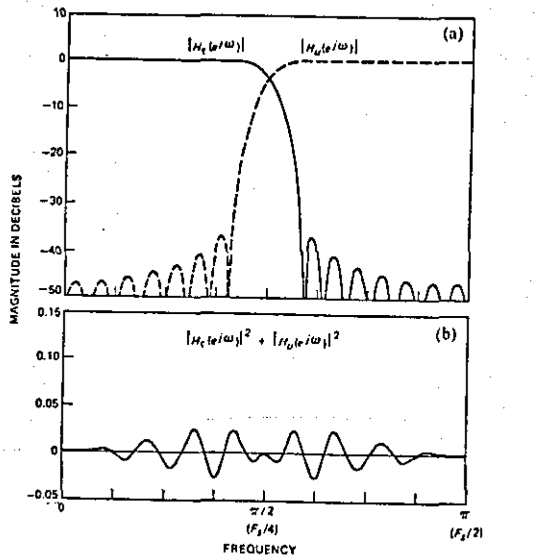


Figure 3.2–43 [88] (a) General block diagram of a two-band sub-band coder. (b) A spectral description of the sub-bands.



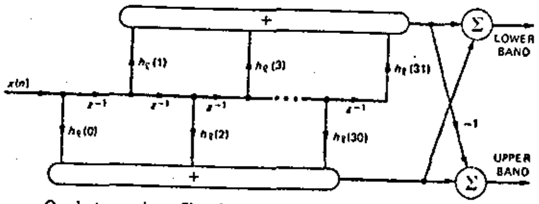
Frequency response for a 32-tap quadrature mirror filter design. (a) Magnitude responses of the individual filters. (b) Magnitude response of the composite system.

Figure 3.2–44 [88]

Quadrature Mirror Filters of order $N = 32$ and $N = 16$. Listed numbers are values of coefficients $h_l(n)$ for $N/2 \leqslant n \leqslant N$. Values of other coefficients follow (11.8) and (11.9)

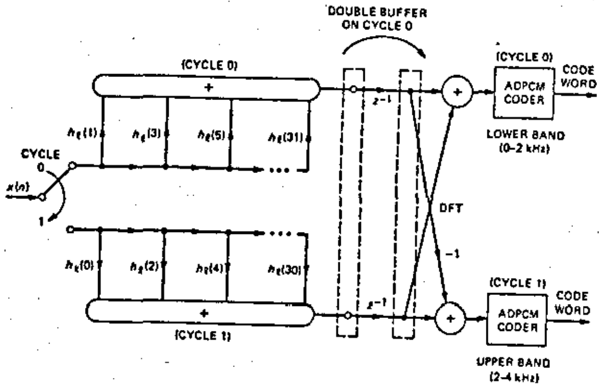
$N = 32$		$N = 16$
$h(16)$ to $h(23)$	$h(24)$ to $h(31)$	$h(8)$ to $h(15)$
4.6645830E-01	1.7881950E-02	.47211220E 00
1.2846510E-01	−1.7219030E-04	.11786660E 00
−9.9800110E-02	−9.3636330E-03	−.99295500E-01
−3.9244910E-02	1.4272050E-03	−.26275600E-01
5.2909100E-02	4.1581240E-03	.46476840E-01
1.4468810E-02	−1.2601150E-03	.19911500E-02
−3.1155320E-02	−1.3508480E-03	−.20487510E-01
−4.1094160E-03	6.5064660E-04	.65256660E-02

Table 3.2–T7 [21]



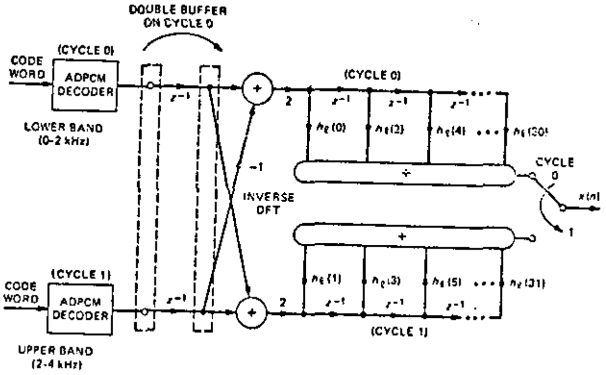
Quadrature mirror filter bank structure that shares computation between upper and lower filters.

Figure 3.2–45 [88]



Sub-band coding transmitter structure using a polyphase QMFB.

Figure 3.2–46 [88]



Sub-band coding receiver structure using a polyphase QMFB for synthesis.

Figure 3.2–47 [88]

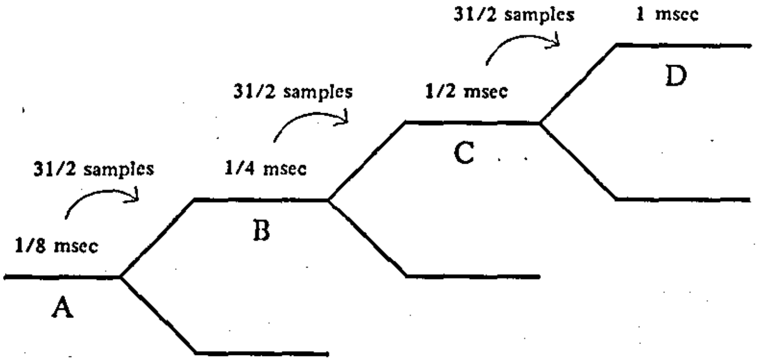


Figure 3.2-48 Filter delay in a three stage tree structured QMF filterbank. The input sample spacing is taken to be 1/8 msec and the sampling frequency is halved at each stage. The number of taps is assumed to be 32 for each filter stage. The delay introduced by each filter is equal to (32–1)/2 samples at the input (to each stage) sampling rate. For example, to calculate the delay between signals B and D proceed as follows : Signal C is delayed 31/2 samples of spacing 1/4 msec w.r.t signal B. Signal D is delayed 31/2 samples of spacing 1/2 msec w.r.t. C. Therefore, the total delay between signals B and D is (31/2)(1/4) + (31/2)(1/2) = (31/2)(3/4) msec. This is equivalent to 3(31/2) = 46.5 samples of signal B (1/4 msec apart).

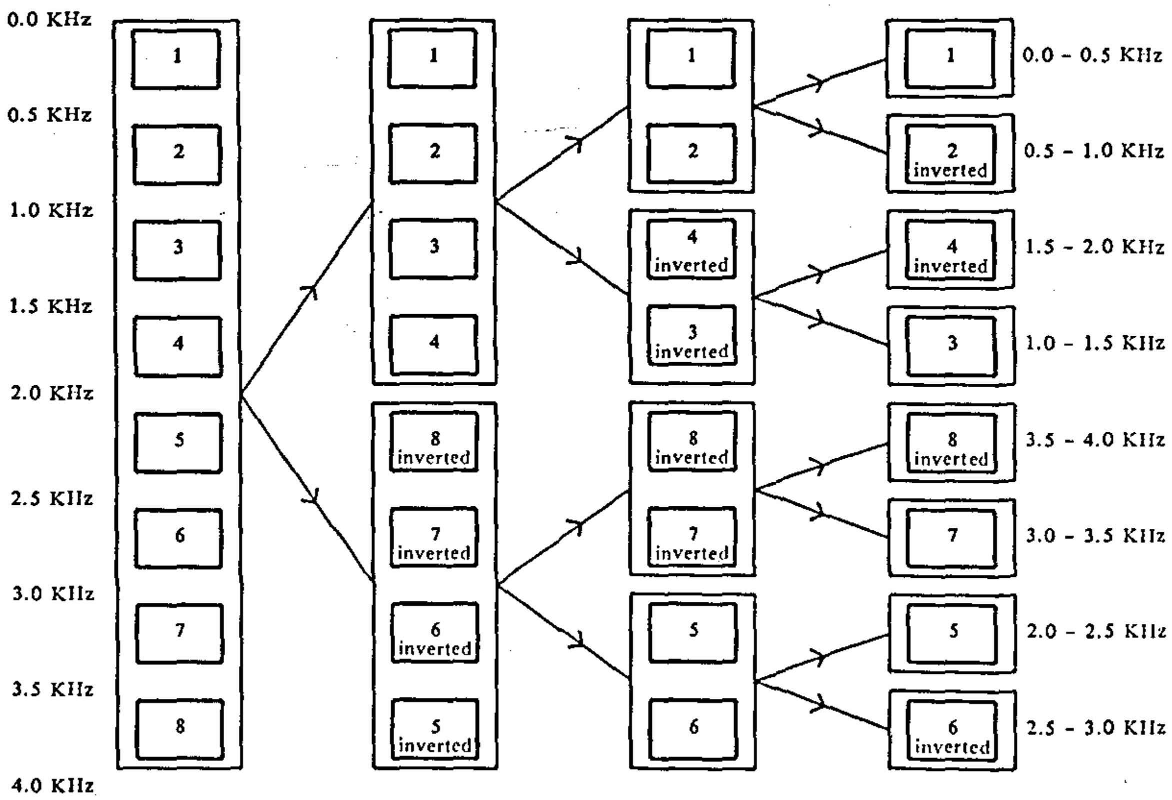
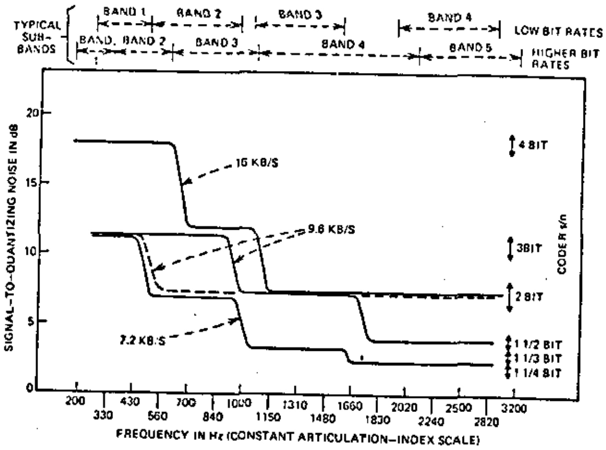
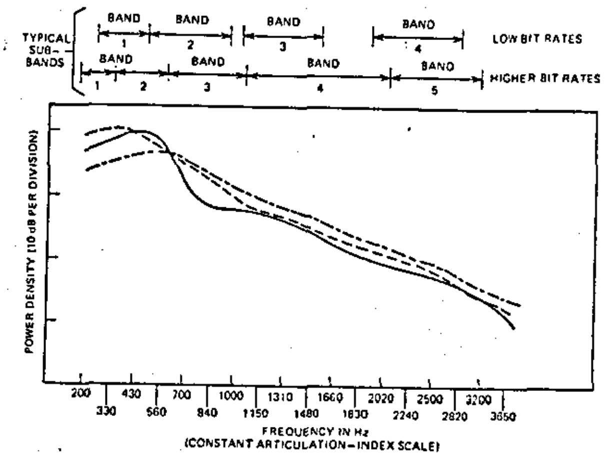


Figure 3.2-49 . Frequency inversion in a three stage, tree structured, QMF analysis bank.



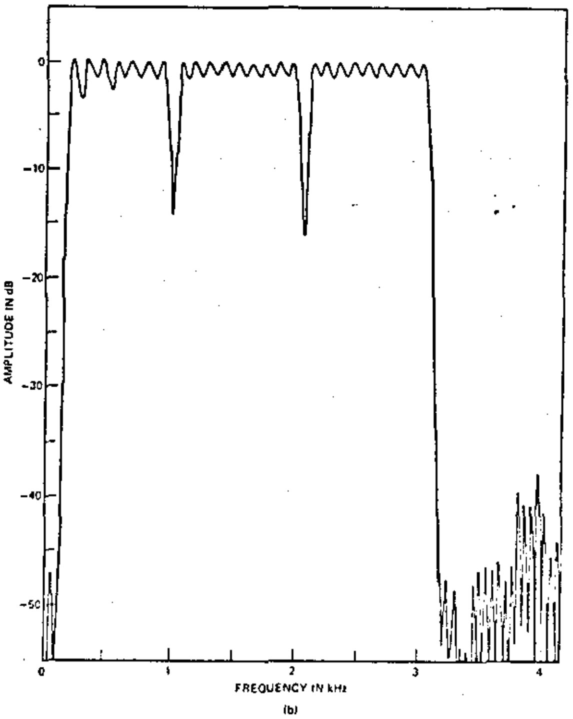
Signal-to-quantizing noise ratio (s/n) as a function of frequency for bit allocations for 16-μ, 9.6-μ, and 7.2-kb/s coders.

Figure 3.2–50 [87]



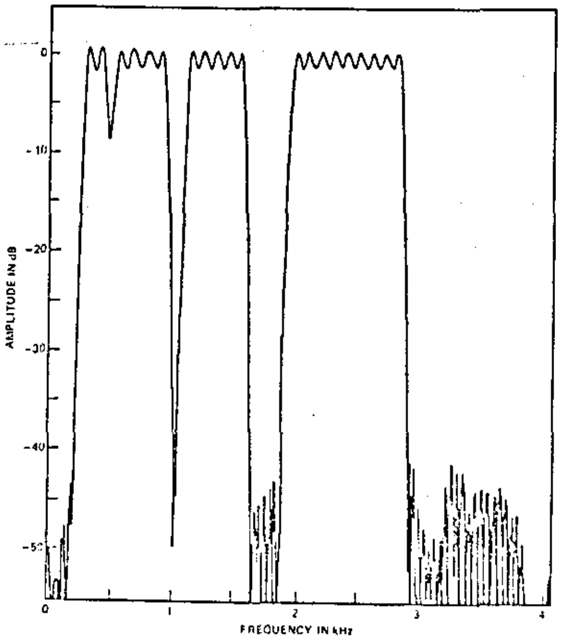
Long-term spectrum of speech Frequency scale based on a constant contribution to the articulation index.

Figure 3.2–51 [87]



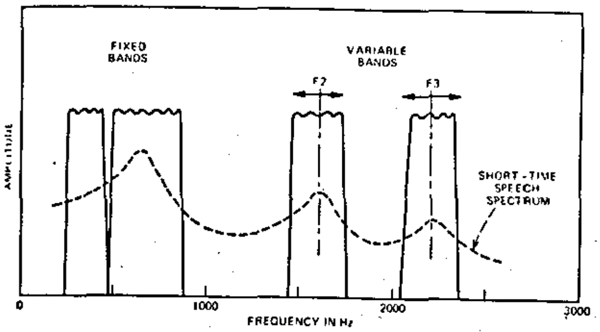
(b)—Measured frequency response for 16-kb/s coder.

Figure 3.2–52b [87]



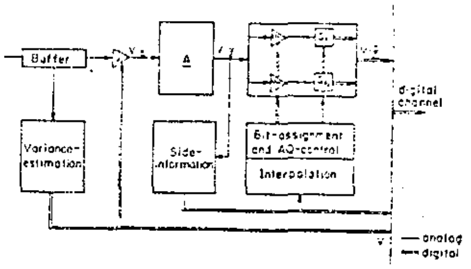
(a)—Measured frequency responses for 7.2- and 9.6-kb/s coders.

Figure 3.2–52a [S7]



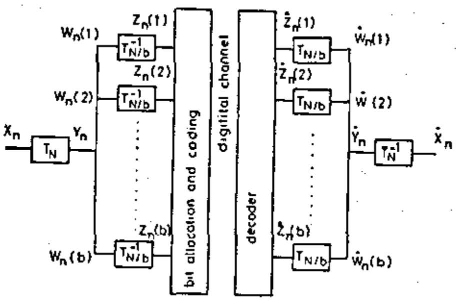
Frequency domain interpretation of the variable-band coder.

Figure 3.2–53 [89]



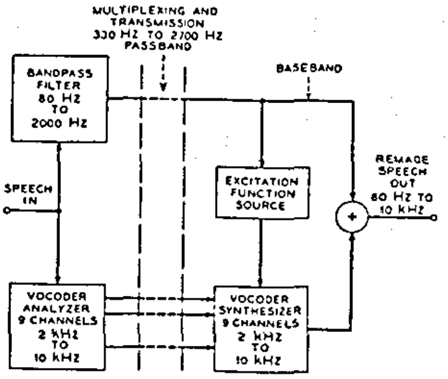
Structure of an ATC-AQF coder. (Characters with underbars are boldface in text.)

Figure 3.2–54 [113]



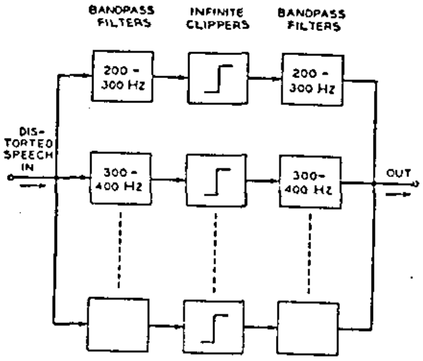
Split-band coder structure (TSBC)

Figure 3.2–55 [115]


MULTIPLEXING AND TRANSMISSION 330 HZ TO 2700 HZ PASSBAND
BANDPASS FILTER 80 HZ TO 2000 HZ
BASEBAND
SPEECH IN
EXCITATION FUNCTION SOURCE
REMADE SPEECH OUT 80 HZ TO 10 KHZ
+
VOCODER ANALYZER 9 CHANNELS 2 KHZ TO 10 KHZ
VOCODER SYNTHESIZER 9 CHANNELS 2 KHZ TO 10 KHZ

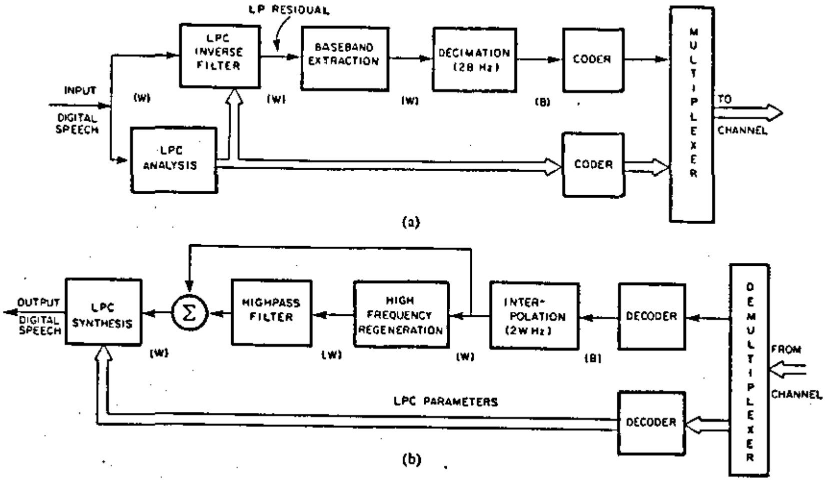
Voice-excited vocoder for transmitting 10-kHz speech signals over 3-kHz telephone lines. The excitation signal at the synthesizer is derived from an uncoded baseband signal by means of a nonlinear "spectrum flattening" process.

Figure 3.2–56 [16]



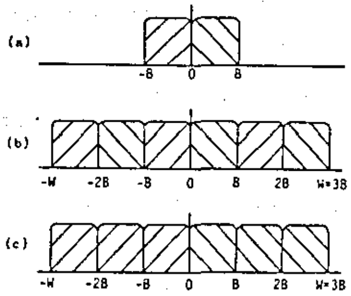
Block diagram of spectrum flattener employing band-pass filters and hard limiters ("infinite clippers").

Figure 3.2–57 [16]



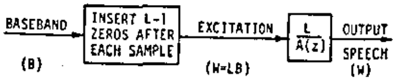
Block diagram of a baseband residual coder. (a) Transmitter. (b) Receiver.

Figure 3.2–58 [116]



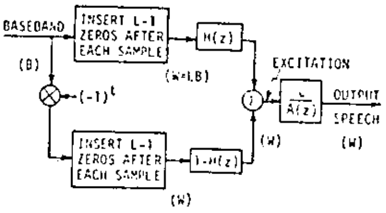
(a) Baseband spectrum, (b) three-band spectral folding, (c) three-band spectral translation.

Figure 3.2–59 [120]



Receiver for baseband coder that uses integer-band spectral folding.

Figure 3.2–60 [120]



Receiver for baseband coder that uses integer-band spectral translation.

Figure 3.2–61 [120]

1. D. O'Shaughnessy "Speech Communication: Human and Machine" ADDISON-WESLEY Publishing Company 1987.

2. L.R. Rabiner, R.W. Schafer "Digital Processing of Speech Signals" Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1976.

3. G. Peterson, & H. Barney (1952). "Control methods used in a study of vowels" J. Acoust. Soc. Am. Vol. 24, pp175-184.

4. A. Holbrook and G. Fairbanks "Diphthong Formants and their Movement" J. of Speech and Hearing Research, Vol. 5, No 1, p38-58 March 1962.

5. J.D. Markel, A.H. Gray, Jr. "Linear Prediction of Speech" Springer-Verlag Heidelberg New York, 1976.

6. B.S. Atal and S.L. Hanauer, "Speech analysis and Synthesis by Linear prediction of the Speech Wave" J. Acoust. Soc. Am. Vol 50, No. 2 (Part 2) pp637-655 August 1971.

7. Atal, B.S. "Determination of the Vocal Tract Shape Directly from the Speech Wave. J. Acoust. Soc. Am. 47, 65(A) 1970.

8. Wakita, H. "Direct estimation of the Vocal tract Shape by inverse filtering of acoustic Speech waveforms" IEEE transactions AU-21 pp417-427, 1973.

9. J.L. Flanagan, M.R. Schroeder, B.S. Atal, R.E. Crochiere, N.S. Jayant, J.M. Tribolet "Speech Coding" IEEE Trans. on Communications, Vol. COM-27, No. 4, April 1979 pp710-736.

10. B.S. Atal "High Quality Speech at low bit rates: Multipulse and Stochastically Excited Linear predictive Coders" Proc. ICASSP 86 pp1681-1684, 1986.

11. J.L. Flanagan, "Speech Analysis Synthesis and Perception" Springer-Verlag - Berlin - Heidelberg, New York 1972.

12. B. Gold, P. Blankenship & R. McAnlay "New applications of channel vocoders" IEEE Trans. ASSP, ASSP-29, pp13-23, 1981.

13. V. Viswanathan, J. Makhoul, R. Schwartz and A. Huggins "Variable frame rate transmission: A review of methodology and application to narrow-band - LPC speech coding" IEEE Trans. Comm. COM-30, 674-686.

14. J. Makhoul, R. Viswanathan, R. Schartz and A.W.F. Huggins "A mixed source model for speech compression and synthesis" IEEE, ICASSP 1978 pp163-166.

15. S. Maitra, C.R. Davis "Improvements in the classical model for beter speech quality" IEEE, ICASSP 1980 pp23-27.

16. M.R. Schroeder "Vocoders: Analysis and Synthesis" Proc. IEEE, Vol. 54, pp720-734, May 1966.

17. A. Oppenheim "Speech analysis-synthesis system based on homomorphic filtering" J. Acoust. Soc. Am. 45, pp459-462.

18. J. Lim "Spectral root homomorphic deconvolution system" IEEE Trans. ASSP, ASSP-27, pp223-233, 1979.

19 G.H. Coker "Computer Simulated Analyser for a formant Vocoder" J. Acoust. Soc. Am. Vol. 35, 1963.

20. S. Roucos, J. Makhoul, R. Schwartz "A variable-order Markov chain for coding of speech spectra" IEEE, ICASSP 1982, pp582-585.

21. N.S. Jayant, P. Noll "Digital Coding of Waveforms" Prentice-Hall Inc. Englewood Cliffs, New Jersey, 1984.

22. S.P. Lloyd "Least Squares Quantization in PCM" IEEE trans. on Information Theory pp129-136 March 1982.

23. J. Max "Quantizing for Minimum Distortion" IRE Trans. on Information Theory pp7-12, March 1960.

24. T.P. Barnwell, A.M. Bush, J.B. O'Neal and R.W. Stroh" Adaptive differential PCM Speech Transmission" Report No. RADC-TR-74-177, Rome air Development Centre, July 1974.

25. N.S. Jayant "Adaptive Quantization with one-word Memory" Bell System Tech. Journal Vol. 52 No.7, Sept. 1973. pp1119-1144.

26. P. Cummiskey, N.S. Jayant, J.L. Flanagan "Adaptive Quantization in differential PCM coding of Speech" Bell System Tech. J. pp1105-1118, Sept. 1973.

27. D.J. Goodman, A. Gersho "Theory of an adaptive Quantizer" IEEE Trans. on Communications pp1037-1045, Aug, 1974.

28. D. Mitra and B. Gotz "An adaptive PCM system Designed for noisy Channels and Digital implementations" Bell System Tech. J. pp2727-2763 Sept. 1978.

29. N.S. Jayant "Adaptive Postfiltering of DPCM Speech" Bell System Tech. J. pp707-717, May-June 1981.

30. P. Noll "A comparative Study of Various Schemes for Speech encoding" Bell System Tech. J. Vol. 54, No. 9, pp1597-1614, Nov. 1975.

31. J. Makhoul "Linear Prediction: A tutorial Review" Proc. IEEE, Vol. 63, pp561-580, 1975.

32. Pfeifer, L.L. "Multiplication Reduction in Short-Term Autocorrelation, IEEE Trans, AU-21, 556-558 (1973).

33. Blankinship, W.A. "Note on Computing Autocorrelation" IEEE Trans. ASSP-22, 76-77, 1974.

34. V. Ramamoorthy and N.S. Jayant "Enhancement of ADPCM speech by adaptive postfiltering" Bell Syst. Tech. J. 63(8), pp1465-1475 (Oct. 1984).

35. B.S. Atal and M.R. Schroeder "Predictive Coding of Speech signls and Subjective Error Criteria" IEEE trans on Acoustics, Speech and Signal Proc. Vol. ASSP-27 No. 3 1979 pp247-254.

36. B.S. Atal "Predictive Coding of Speech at low bit rates" IEEE trans on Communications Vol-COM-30, No. 4 April 1982. pp600-614.

37. M. Krasner, M. Berouti, J. Makhoul "Stability Analysis of APC Systems" IEEE, ICASSP 1981, pp627-630.

38. S. Singhal and B.S. Atal "Improving Performance of Multipulse LPC coders at low bit rates" IEEE, ICASSP 1984 pp1.3.1-1.3.4

39. J. Makhoul "Stable and Efficient Lattice methods for Linear Prediction" IEEE Trans Acoust. Speech and Signal Proc. Vol. ASSP-25, No. 5, pp423-428, Oct. 1977.

40. J. Burg "A new analysis Technique for Time Series Data" Proc. NATO Advanced study Institute on Signal Proc. Enschede Netherlands, 1968.

41. G. Rebolledo, R.M. Gray, J.P. Burg "A Multirate Voice Digitizer based upon Vector Quantization" IEEE Trans. on Communications Vol. COM-30, No. 4 - April 1982 pp721-727.

42. J. Makhoul "Specral Analysis of Speech by Linear Predictiion" IEEE Trans. on Audio and Electroacoustics Vol AU-21, No 3, pp140-148, June 1973.

43. J. Makhoul "Methods for nonlinear spectral distortion of speech signals" Proc. 1976 Int. Conf. Acoust. Speech Signal Proc. (Philadelphia) 87-90.

44. H.W. Strube "Linear Prediction on a Warped frequency Scale" J. Acoust. Soc. Am. 68(4) Oct. 1980 pp1071-1076

45. J. Makhoul, L. Cosell "LPCW: An LPC Vocoder with Linear Predictive Spectral Warping" Proc. 1976. ICASSP (Philadelphia) pp466-469.

46. J.E. Roberts and R.E. Wiggins "Piecewise Linear Predictive Coding (PLPC)" ICASSP 1976 pp470-473.

47. H. Hermansky, H. Fujisaki, Y. Sato "Analysis and Synthesis of Speech based on Spectral Transform Linear Predictive method" IEEE, ICASSP 1983 (BOSTON) pp777-780.

48. H. Hermansky, B.A. Hanson and H. Wakita "Perceptually based Linear Predictive analysis of Speech" IEEE ICASSP 1985 pp509-512 (13-10.1 13.10-4).

49. M.R. Scroeder "Predictive Coding of Speech: Historical Review and directions for future research" IEEE, ICASSP 1986 (Tokyo) pp3157-3164.

50. R. Viswanathan, J. Makhoul "Quantization Properties of Transmission Parameters in Linear Predictive Systems" IEEE Trans on Acoustics, Speech and Signal Proc. Vol. ASSP-23 No 3, June 1975 pp309-321.

51. A.H. Gray, Jr, and J.D. Markel "Distance measures for Speech Processing" IEEE Trans. on Acoustics, Speech and Signal Processing, Vol ASSP-24 No 5, Oct. 1976 pp380-391.

52. J. Makhoul, S. Roucos, H. Gish "Vector Quantization in Speech coding" IEEE proc. Vol. 73, No 11, Nov. 1985. pp1551-1588.

53. Y. Linde, A. Buzo, R.M. Gray "An algorithm for Vector Quantizer Design" IEEE Trans. on Comm. Vol. COM-28, No 1, Jan, 1980.

54. A. Buzo, A.H. Gray Jr., R.M. Gray, J.D. Markel "Speech Coding based upon Vector Quantization" IEEE Trans on Acoustics Speech and Signal Proc. Vol. ASSP-28 No 5, Oct. 1980.

55. R.M. Gray, H.Abut "Full search and tree search Vector Quantization of Speech waveforms" IEEE, ICASSP 1982, p593-596.

56. M. Copperi and D. Sereno "9.6kbits/s piecewise LPC residual excited coder using multiple-stage vector quantization" Proc IEEE Int. Conf. Acoust, Speech, Signal Proc. (San Diego, CA, Mar. 1984) paper 10.5.

57. M.J. Sabin and R.M. Gray "Product Code Vector Quantizers for waveform and Voice Coding" IEEE Trans on Acoust. Speech and Signal Proc. Vol. ASSP-32, No 3, June 1984. pp474-488.

58. M. Copperi and D. Sereno "Feature exraction and product Codes in vector excited coders" Proc IEEE, ICASSP 1987 pp1942-1945.

59. T. Moriya and M. Honda "Transform coding of speech with weighted Vector Quantization" Proc. IEEE, ICASSP 1987 pp1629-1632.

60. S. Andlersberg and V. Cuperman "Transform domain Vector Quantization for Speech Signals" Proc IEEE ICASSP 1987 pp1938-1941.

61. S. Singhal "On encoding filter parameters for stochastic coders" Proc. IEEE, ICASSP 1987 pp1633-1636.

62. B.S. Atal "Stochastic Gaussian Model for Low-Bit Rate Coding of LPC Area Parameters" ICASSP 1987 pp2404-2407.

63. M.O. Dunham and R.M. Gray "An algorithm for the Design of Labelled-Transition Finite-State Vector Quantizers" IEEE Trans on Comms. Vol COM-33 No 1. Jan 1985 pp83-89.

64. Y. Shoham "Vector Predictive Quantization of the Spectral Parameters for low rate speech coding" ICASSP 1987 pp2181-2184.

65. S. Roucos, R. Schwartz and J. Makhoul "Segment Quantization for very low rate speech coding" Proc IEEE ICASSP 1982 pp1565-1569.

66. J. Makhoul and M. Berouti "Adaptive noise spectral shaping and entropy coding in predictive coding of speech" IEEE Trans, Acoust, Speech, Signal Proc. Vol ASSP-27 pp63-73, Feb. 1979.

67. B.S. Atal and M.R. Schroeder "Optimizing Predictive Coders for Minimum Audible Noise" Proc. IEEE, ICASSP 1979 pp453-455.

68. B.J. McDermott, C. Scagliola "The Perception of Spectrally Shaped Additive Noise in Speech" IEEE Proc. ICASSP 1982, pp196-198.

69. M.R. Schroeder, B.S. Atal, J.L. Hall "Optimizing digital speech coders by exploiding masking properties of the human ear" J. Acoust. Soc. Am., 66(6), Dec. 1979, pp1647-1652.

70. H.G. Fehn and P. Noll "Multipath Search Coding of Stationary Signals with Applications to Speech" IEEE Trans. Communications Vol-COM-30 No 4, April 1982 pp687-701.

71. J.H. Chen and A. Gersho "Vector Adaptive Predictive Coding of Speech at 9.6kb/s" IEEE Proc. ICASSP 1986 pp1693-1696.

72. J.H. Chen and A. Gersho "Real-time Vector APC Speech coding at 4.8 Kb/s with adaptive Postfiltering" IEEE Proc. ICASSP 1987 pp2185-2188.

73. B.S. Atal and M.R. Schroeder "Stochastic coding of speech signals at very low bit rates" Proc. IEEE Int. Conf. Communications p48.1 1984.

74. M.R. Schroeder and B.S. Atal "Code-Excited Linear prediction (CELP): High qualty speech at very low bit rates" Proc. IEEE Int. Conf. Acoust. Speech, Signal Proc. pp937-940, 1985.

75. J.P. Adoul and C. Lamblin "A comparison of some algebraic structures for CELP coding of speech" Proc. IEEE ICASSP 1987 pp1953-1956.

76. J.P. Adoul, P. Mabilleau, M. Delprat and S. Morissette "Fast CELP coding bsed on algebraic codes" ICASSP 1987 pp1957-1960.

77. M.R. Schroeder, B.S. Atal "Speech Coding Using efficient block codes" IEEE Proc ICASSP 1982 pp1668-1671.

78. M.R. Schroeder and B.S. Atal "Rate distortion theory and predictive coding" IEEE Proc. ICASSP 1981 pp201-204.

79. B.S. Atal and J.R. Remde "A new model for LPC excitation for producing natural-sounding speech at low bit rates" Proc. 1982, ICASSP pp614-617.

80. T. Araseki, K. Ozawa, S. Ono and K. Ochiai "Multi-pulse excited speech coder based on maximum crosscorrelation search algorithm" IEEE Proc. Global Telecom. Conf. 1983, pp794-798.

81. P. Kroon, E.F. Deprettere, R.J. Sluyter "Regular-Pulse excitation - A novel approach to effective and efficient Multipulse Coding of Speech" IEEE Trans ASSP Vol. ASSP 34 No 5, Oct 1986 pp1054-1063.

82. P. Kroon and E.F. Deprettere "Experimental evaluation of different approaches to the multipulse coder" ICASSP 1984 pp10.4.1-10.4.4

83. K. Ozawa, T. Araseki "High Quality Multipulse speech coder with pitch prediction" IEEE Proc. ICASSP 1986 pp1689-1692.

84. K. Ozawa, T. Araseki "Low bit rate Multi-pulse Speech Coder with Natural Speech Quality" IEEE Proc. ICASSP 1986 pp457-460.

85. J.M. Tribolet, R.E. Crochiere "Frequency Domain coding of Speech" IEEE Trans Acoustics Speech, Signal Proc. Vol ASSP-27, No 5, Oct. 1979, pp512-530.

86. R.E. Crochiere, S.A. Webber, J.L. Flanagan "Digital Coding of speech into sub-bands" Bell System Tech. Journal Vol, 55, No 8, pp1069-1085, Oct. 1976.

87. R.E. Crochiere "On the design of sub-band coders for low-bit-rate speech communication" Bell Sys. Tech. Jour. Vol. 56, pp747,770, 1977.

88. R.E. Crochiere "Sub-band Coding" Bell Sys. Tech. Jour. Vol.60, pp1633-1654, 1981.

89. R.E. Crochiere and M.R. Sambur "A Variable-band Coding Scheme for Speech Encoding at 4.8 kb/s" Bell. Syst. Tech. Jour. Vol. 56, No 5, pp771-780, 1977.

90. R.E. Crochiere and L.R. Rabiner "Interpolation and decimation of digital signals - a tutorial review" Proc. IEEE, Vol, 69, No 3, pp300-331, 1981.

91. D. Esteban and C. Galand "Application of Quadrature mirror filters to split band voice coding schemes" Conf. Proc. ICASSP 1977 pp191-195.

92. J.D. Johnston "A filter family designed for use in Quadrature mirror filter banks" Proc IEEE ICASSP 1980 ppp291-294.

93. C.R. Galand and D.J. Esteban "Design and evaluation of Parallel Quadrature mirror filters (PQMF)" Proc. IEEE ICASSP 1983, pp224-227.

94. R.A. McDonald and P.M. Schultheiss "Information rates of Gaussian signals under criteria constraining the error spectrum" Proc. IEEE Vol 52, pp415-416, 1964.

95. T. Ramstad "Subband Coder with a simplified Adaptive bit alocation algorithm" IEEE Conf. Proc. ICASSP 1982, pp203-207.

96. J.M. Tribolet, P. Noll, B.J. McDermott, R.e. Crochiere "A study of complexity and Quality of speech waveform coders" Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc pp586-590 1978.

97. J.M. Tribolet, P. Noll, B.J. McDermott and R.E. Crochiere "A Comparison of the prformance of four low bit rate speech waveform coders" BSTJ, pp699-713, March 1979

98. R.V. Cox "A comparison of three speech coders to be implemented on the digital signal processor" BSTJ Vol 60, pp1411-1421, 1981.

99. V. Gupta, K. Virnpaksha "Performance evaluation of adaptive quantizers for a 16 kb/s sub-band coder", ICASSP 1982 pp1688-1691.

100. R.E. Crochiere, R.V. Cox, J.D. Johnston "Real time speech coding" IEEE Trans. Commun. vol COM-30 No 4, pp621-633 1982.

101. R.E. Crochiere "A novel Approach for implementing Pitch prediction in subband Coding" ICASSP 1979 pp526-529.

102. V.R. Daumer, "Subjective evaluation of several efficient speech coders" IEEE Trans. Commun Vol. COM-30 No 4, pp655-662, 1982.

103. B.S. Atal and J.R. Remde "Split band APC system for low bit rate encoding of speech" Proc. ICASSP, pp599-601, April 1981.

104. M. Honda, F. Itakura "Bit allocation in time and frequency domains for predictive coding of speech" IEEE Trans Acoust. Speech Signal Ptoc. Vol. ASSP-32 No 3, pp465-473, June 1984.

105. T.A. Ramstad, O. Foss" Subband Coder design using recursive Quandrature mirror filters Proc. EUSIPCO, pp747-752., 1980

106. H. Abut, S.A. Luse "Vector Quantizers for subband coded wabeforms" Proc IEEE, Int. Conf. Acoust, Speech,Signal Proc. pp10.6.1-10.6.4, 1984.

107. A. Gersho, T. Ramstad, I. Versvik "Fully vector-quantized subband coding with adaptive codebook allocation" Proc. IEEE Int. Conf. Acoust. Sp. Sig. Proc pp10.7.1-10.7.4 1984.

108. I. Versik, H.C. Guren "Subband coding with Vector Quantization" Proc. IEEE ICASSP 1986 pp3099-3102.

109. N. Ahmed, T. Natarajan, K.R. Rao "Discrete Cosine Transform" IEEE Trans on Computers Jan 1974 pp90-93.

110. W.H. Chen, C.H. Smith, S.C. Fralick "A fast computational Algorithm for the Discrete Cosine Transform" IEEE Trans on Comm. Vol. COM-25 No 9, Sept. 1977 pp1004-1009.

111. J. Makhoul "A fast Cosine Transform in one and two dimensions" IEEE Trans, ASSP, Vol-ASSP-28 No 1 Feb. 1980 pp27-34.

112. G. Bertocci, B.W. Schoenherr, D.G. Messerschmitt "An approach to the implementation of a Discrete Cosine Transform" IEEE Trans on Commun, Vol. COM-30 No 4 April 1982 pp635-641.

113. R. Zelinski, P. Noll "Adaptive Transform Coding of Speech Signals" IEEE Trans ASSP, Vol. ASSP-25, No 4, 1977, pp299-309.

114. R. Zelinski, P. Noll "Approaches to Adaptive Transform Speech Coding at low bit rates" IEEE Trans ASSP Vol-ASSP-27, No. 1, Feb. 1974 pp89-95.

115. F.S. Yeoh, C.S. Xydeas "Transform approach to split-band coding schemes "IEE proc. Vol. 131, part f No 1, 1984, pp57-63.

116. V.R. Viswanathan, A.L. Higgins, W.H. Russell "Design of a Robust Baseband LPC coder for Speech Transmission Over 9.6 kbits/sec Noisy Channels" IEEE Trans on Comm. Vol. COM-30, No 4, April 1982, pp663-673.

117. B.S. Atal, M.R. Schroeder, V.Stover "Voice-excited predictive coding system for low bit-rate transmission of speech" Int. Conf. Commun. San Francisco, 16-18 June, 1975 pp30.37-30.40

118. C. Galand, K. Daulasim, D. Esteban "Adaptive Predictive Coding of base-band speech signals" IEEE proc. ICASSP 1982 pp220-222.

119. H. Katterfelt "A DFT-based residual-excited Linear Predictive coder (RELP) for 4.8 and 9.6 kb/s" IEEE Proc ICASSP 1981 pp824-827.

120. J. Makhoul and M. Berouti "Predictive and Residual Encoding of Speech" J. Acoust. Soc. Am. 1979 pp1633-1641.

121. C.K. Un and J.R. lee "On spectral Flattening Techniques in Residual-Excited Linear prediction Vocoding" IEEE proc ICASSP 1982, pp216-219.

122. R.J. Sluyter, G. J, Bosscha, H.M.P.T. Schmitz "A 9.6 kbits/s speech coder for mobile radio Applications" in "Links for the future" Science, Systems and Services for Communications P. Dewilde and C.A. May (editors) IEEE/Elsevier Science Publishers B.V (North-Holland), 1984.

CHAPTER 4

DISTORTION, DISTORTION MEASUREMENTS

CHAPTER 4
DISTORTION, DISTORTION MEASUREMENTS

4.1 Rate Distortion Theory

The mean square error (mse) is a widely used distortion criterion. Its analytical tractability has undoubtedly contributed to its widespread acceptance. However, its ability to provide a meaningful measure of speech quality should not be underestimated. In many situations of waveform coding the mse reflects the true perceptual impact of rate compression and in others it provides a building block from which more appropriate measures can be constructed.

Rate distortion theory has generally built upon the mse [1], although other distortion measures can be used to derive rate distortion functions [2].

The following discussion assumes stationary sources and therefore any results must be applied in a "piecewize" manner in the case of speech coding and for time durations for which the speech signal can be considered as stationary.

A rate distortion function D(R) relates the minimum attainable distortion D (which we assume to be represented by the reconstruction error variance) to the bit rate R, for a particular source.

A closely related function R(D) relates the minimum bit rate required to encode the source with a distortion D.

For a memoryless (i.e. flat spectrum, white) zero-mean Gaussian source of variance σ_x^2 the above two functions are given by:

$$R(D_w)=\max\left\{0,\frac{1}{2}\log_2\frac{\sigma_x^2}{D_w}\right\} = \begin{cases} \frac{1}{2}\log_2\frac{\sigma_x^2}{D_w}; & 0\leqslant D_w\leqslant\sigma_x^2 \\ 0; & D_w\geqslant\sigma_x^2 \end{cases} \qquad 4.1a$$

$$D_w(R) = 2^{-2R}\sigma_x^2 \qquad 4.1b$$

In the case of $D_w \geqslant \sigma_x^2$ no information needs to be transmitted since by using zero samples for reconstruction a distortion equal to σ_x^2 will be obtained. For a coder operating at the rate distortion bound, the output Y to a zero mean gaussian input x is given by [1,3]

$$y = \left(1 - \frac{D_w}{\sigma_x^2}\right).x + n \qquad 4.2a$$

with n being a zero mean gaussian x-independent additive noise of variance

$$\sigma_n^2 = D_w\left(1 - \frac{D_w}{\sigma_x^2}\right) \qquad 4.2b$$

which ensures that the variance of the reconstruction error r = y - x is equal to D_w. Note that equations 4.2a,b are essential for $\sigma_r^2 = D_w$ to be true, where D_w is the minimum attainable distortion at the given bit rate. Instantaneous quantizers cannot provide the conditions described by 4.2 and are therefore suboptimal. Delayed decision coders on the other hand strive to achieve relations 4.2a,b and thus provide a performance closer to the rate distortion bound.

Sources with memory (i.e. a non-flat spectrum) permit greater data compression than memoryless sources for a given D. In fact the D(R) function reflects this higher data compression through expressions that measure the non-flatness of the source's power spectral density (psd).

The relationship between higher compression and non-flatness can be easily visualized in the case of a single peak in the psd: The spectral peak implies a higher presence of those frequency components in the signal which constitute the peak. This in turn implies a certain degree of predictability about the signal. It is this predictability which reduces the amount of information needed for signal reconstruction. We have seen that predictive, transform and subband coders exploit this predictability to approach the rate distortion bound for sources with memory such as speech. The nonstationarity of the speech signal is the factor that necessitates the transmission of side information in the above coders.

Let the psd of the input be given by $S_x(f)$. The D(R) function for a general (non flat, coloured) zero-mean Gaussian source is given parametrically in the form [1,3,4,5]

$$D(A) = \frac{1}{F} \int_{o}^{F} \min [A, S_x(f)] df \qquad 4.3a$$

$$R(A) = \frac{1}{F} \int_{o}^{F} \max [0, \frac{1}{2} \log_2 \frac{S_x(f)}{A}] df \qquad 4.3b$$

For a minimum reconstruction error psd N(f), the area E shown in figure 4.1 is equal to D(A). The frequency axis is divided into a number of passbands ($f \epsilon \alpha$) and stopbands ($f \epsilon \beta$). The total area E of the noise spectrum "fills up" part of the area under the signal spectrum much as a liquid fills an irregular container (this is the so-called "water-filling" procedure).

Frequency ranges which are completely "filled up" make no contribution to the information rate. Therefore rate distortion theory dictates some form of noise shaping, although the level in the passbands is flat. Equations 4.3a,b can only be satisfied through delayed decision coding procedures. One such procedure is depicted in figure 4.2. A(Z) is the optimum linear predictor for Y_n, σ^2 the mean squared prediction error. The sequence $\hat{V}_n$ is restricted to be a zero-mean unit variance white noise process. This necessitates the introduction of filter C(Z) defined by

$$|C(z)|^2 = \max [0, 1- \frac{A}{\sigma^2} |1-A(Z)|^2] \qquad 4.4$$

(A is the parameter in 4.3)

so that equations 4.3a,b are satisfied. [5] Note that in coding algorithms such as multipulse LPC where the mse is minimized explicitly, C(Z) can perhaps be omitted. In effect, the generated sequence $\hat{V}_n$ is already (approximately) shaped by an "internal" filter C(Z). The shaping results from explicitly minimizing the mse. The above can be easily modified to derive rate distortion functions where the criterion is a weighted mse [3,4,5]. In particular the problem of encoding a source with spectrum $S_x(f)$ and an error spectrum AE(f) is equivalent to the problem of coding a source with spectrum S(f)/E(f) and a flat error spectrum A. Going back to equation 4.3, for sufficiently low distortion (i.e. a high enough bit rate) such that $A \leq \min\{S_x(f)\}$, then D(A) = A and

$$R(A) = R(D) = \frac{1}{F} \int_0^F \frac{1}{2} \log_2 [\frac{S_x(f)}{D}] \, df \qquad 4.5$$

By equating R(D) from above to R(D$_\omega$) of 4.1, it is straightforward to show that (appendix E):

$$10\log_{10}\left(\frac{D}{D_w}\right)=10\log_{10}\left[\frac{1}{F}\int_{o}^{F} S_x(f)df\right]$$

$$-\frac{1}{F}\int_{o}^{F} 10\log_{10}S_x(f)df \qquad 4.6$$

(subject to the condition that $\int_{o}^{F} S_x(f)df$ is the same in both cases) which, is exactly the limiting prediction gain for a nonuniform spectrum (e.g. equations 3.2-43 and 3.2-44).

It is important to realize that the value given by equation 4.6 for the prediction gain can only be attained under the assumptions of equation 4.5 i.e. $A \leqslant \min(S_x(f))$. For higher distortions (i.e. lower bit rates) the prediction gain is reduced accordingly.

4.2 OBJECTIVE DISTORTION MEASURES

It is often necessary to be able to compare the performance of speech coders in terms of speech quality attained. Ideally the comparison must be made subjectively and under the same conditions for the coders involved. Furthermore the tests should ensure that the ensemble of speech material used in the tests, and the observers judging the coders are representative. Such subjective tests are time-consuming and expensive to implement. Therefore, various objective measures have been devised whose results are correlated with subjective results. The simplest measure that has been used is the mean square error (mse). For the coder shown in figure 4.3 this measure is given by:

$$\sum_{n} r^2(n) = \sum_{n} [Y(n) - X(n)]^2 \qquad 4.7$$

The summation is carried out over the whole of the speech segment. Since the input X(n) can be amplified or attenuated by a large factor without any effect on output speech quality for most coders, a more sensible measure is given by

$$SNR = \frac{\sum_{n} X^2(n)}{\sum_{n} r^2(n)} \qquad 4.7a$$

or in logarithmic form

$$SNR(dB) = 10\log_{10} \frac{\sum_{n} X^2(n)}{\sum_{n} r^2(n)} \qquad 4.7b$$

The summations are again carried out over the whole speech segment.

Since speech is a nonstationary process, coder quality is affected by the distribution of quantization distortions in time. Better indications of perceptual quality can be obtained by using short-time (~ 20 msec) objective measures, "averaged" over a speech utterance, than from conventional long-time SNR. One assumes that the measure given in 4.7, is a valid one for a short speech segment where the summation over n is carried out over about 20 msec were the speech signal is assumed to be stationary. One then obtains a distortion (or equivalently a quality) measure which is a function of time. This is not very useful in itself since what is usually required is a single number representative of the whole speech segment. Therefore it is necessary to know how to "average" the numbers obtained from 4.7 to obtain a single number measure.

This is not a trivial problem since the measure in 4.7 may have a different perceptual equivalent for different short speech segments. This problem of averaging has yet to receive anywhere near the attention it merits and further research is required into this problem. A simple solution is to use the arithmetic mean of 4.7b (or, equivalently, the geometric mean of 4.7a) to obtain the segmental SNR measure:

$$\text{SNRSEG (dB)} = \frac{1}{L} \sum_{L} 10\log_{10} \frac{\sum_{n} X^2(n)}{\sum_{n} r^2(n)} \qquad 4.8$$

where the summation in n is carried out over short segments of duration around 16-20 msec. L is the total number of these segments. The segmental SNR has been shown to be a better measure of coder quality than the SNR as given by 4.7, especially for coders using time-domain or frequency domain prediction. [6,7,8,9]. One problem which can occur with SNRSEG is that in regions of silence the input signal to the coder is essentially zero and any slight noise at the output will give rise to large negative SNRSEG values. These values when averaged together with the segmental values during actual speech, can have a significant effect on the final result. A simple procedure to avoid this problem is to incorporate a threshold test in order to exclude silent segments from the computation [6]. If silent segments are important for the performance of the coder (i.e. in order to assess idle channel coding noise) other modifications can be made [9].

The averaging procedure over time implied by 4.8 will be valid provided the variation of individual measurements around the average is small. For large variations the application of 4.8 is dubious. One generalization of the measure could be the Lp norm defined by

$$Lp = \left[\frac{1}{L} \sum_{L} \left[10\log_{10} \frac{\sum_{n} X^2(n)}{\sum_{n} r^2(n)} \right]^{P} \right]^{1/P} \qquad 4.9$$

As p is varied from 1 to ∞, the contributions to the sum from the larger segmental distances increase. At p=1 these are given uniform weighting irrespective of magnitude. At p=∞ the largest distance is given a weighting of one and all others a zero weighting. The value of p best matching subjective measurements can be chosen. Although the Lp norm provides more flexibility, it does not take into account that the same value of segmental distance between coded and uncoded speech may have a different meaning according to the segment being coded.

We have seen previously that the distortion present at the output of a coder can be modelled by an attenuation term and an uncorrelated component. This relationship is described by equation 4.2. The input for the coder described by equation 4.2 is white. For nonwhite inputs the attenuation component is frequency dependent. The uncorrelated noise component produces distortion which is perceptually more significant than the attentuation term. A better distortion measure could therefore aim to isolate this type of distortion. It is shown [6] that the SNRSEG given by 4.8 produces results that are close to an SNRSEG measure which uses the attenuated signal to uncorrelated noise ratio as its segmental distance. This is one of the reasons why the SNRSEG measure of 4.8 correlates better with subjective performance than the SNR measure given by equation 4.7.

Further refinements to a distortion measure can be conceived if the frequency selectivity of the ear is taken into account. We have seen that the auditory mechanism performs a short-time spectral analysis of the signal at the input to the ear. The nonlinear warping of the frequency scale to the critical band scale can be taken into account to produce a frequency weighted measure. This is also closely related to the classical articulation studies of French and Steinberg [10] and Kryter [11,12]. The SNRSEG measure now becomes segmented in the frequency domain as well, but the frequency bins increase in size with frequency according to the critical band scale or the Articulation Index (A.I.) scale. One such division is shown in table 4.T1. The segmental (in both frequency and time) SNR can then be given by

$$AI = \sum_{i=1}^{20} A_i = 0.05(SNRSEG_i/30) \qquad 4.10$$

where $SNRSEG_i$ is the segmental SNR in band i. The constants in 4.10 relate to the facts that the $SNRSEG_i$ is restricted to a maximum of 30 dB (signifying that a higher value is perceptually unnoticeable) and that a perfect signal is normalized to have an AI of one. The nonlinear division of the frequency bands according to the AI is equivalent to weighting the conventional SNRSEG of equation 4.8 by a frequency dependent factor F(f), shown in figure 4.4 by the solid line. The weighting resulting from a critical band warping is also shown by the dashed line.

Further knowledge about the perception of sounds by the ear can be built in the model by taking into account masking effects and the transformation of intensity into loudness. One such model due to Schroeder et al [13] was presented in the section on psychophysical modelling in the chapter on Hearing.

4.2.1 Spectral Distance Measures

Spectral distance measures are based upon transformations which retain only the smooth spectral behaviour of the speech signal. Although such measures may have limited applicablity for waveform coded speech on their own, they can be used in combination with the waveform orientated measures presented above, to provide a composite measure of quality. In addition their importance has grown in recent years since they provide a basis for vector quantization of the filter parameters in predictive coders. Traditionally such measures have been used in vocoder design, recognition [14] and identification and verification tasks.

We will present spectral distance measures under the assumption that the comparison is made between two all pole spectral models $\sigma/A(Z)$ and $\sigma'/A'(Z)$. The first can be the all pole spectral model of the uncoded speech whereas $\sigma'/A'(Z)$ can be the all pole spectral model of the coded speech. Central to these spectral measures is the log magnitude difference given by [15, 16].

$$V(\theta) = \ln[\sigma^2/|A(e^{j\theta})|^2] - \ln[(\sigma')^2/|A'(e^{j\theta})|^2] \qquad 4.11$$

Most measures can be related to the likelihood ratios which have already been mentioned in the section on vector Quantization of chapter 3. These are defined in terms of δ, α, δ' and α' which are in turn defined by

$$\alpha = \sum_{n=-p}^{P} r_a(n) r_x(n) \qquad 4.12$$

$$\delta = \sum_{n=-p}^{P} r'_a\,(n)\; r_x(n) \qquad 4.13$$

$$\alpha' = \sum_{n=-p}^{P} r_a{}'(n)\; r'_x(n) \qquad 4.14$$

$$\delta' = \sum_{n=-p}^{P} r_a(n)\; r'_x\,(n) \qquad 4.15$$

P is the order of the all-pole filters, $\{r_a(n)\}$ and $\{r'_a(n)\}$ are the autocorrelation sequences of the filter coefficients for the uncoded and coded speech respectively, $\{r_x(n)\}$ and $\{r'_x(n)\}$ are the autocorrelation coefficients of the data $\{X(n)\}$ and $\{X'(n)\}$ respectively, for which $\sigma/A(Z)$ and $\sigma'/(A'(Z)$ represent the optimum models. The terms defined by 4.12-4.15 will only appear in two ratio forms, the likelihood ratios δ/a and δ'/a'. These can be shown to be equal to [15]:

$$\frac{\delta}{\alpha} = 1 + \int_{-\pi}^{+\pi} \frac{|A'(e^{j\theta})-A(e^{j\theta})|^2}{|A(e^{j\theta})|^2} \; \frac{d\theta}{2\pi} \qquad 4.16$$

and

$$\frac{\delta'}{\alpha'} = 1 + \int_{-\pi}^{+\pi} \frac{|A'(e^{j\theta})-A(e^{j\theta})|^2}{|A'(e^{j\theta})|^2} \; \frac{d\theta}{2\pi} \qquad 4.17$$

It can be seen that the differences between the all pole spectra are most heavily weighed when $1/|A(e^{j\theta})|$ (or $1/|A'(e^{j\theta})|$) are large i.e. at the formant peaks.

Most spectral measures are variants of or approximations to the Itakura-Saito measure defined by

$$IS = \int_{-\pi}^{\pi} [e^{V(\theta)} - V(\theta) - 1] \frac{d\theta}{2\pi} \qquad 4.18a$$

or equivalently [15]

$$IS = \left(\frac{\sigma}{\sigma'}\right)^2 \left(\frac{\delta}{\alpha}\right) - 2\ln\left(\frac{\sigma}{\sigma'}\right) - 1 \qquad 4.18b$$

For small $V(\theta)$ the term in the brackets in 4.18a can be approximated by $V^2(\theta)/2$ (by expanding $e^{V(\theta)}$ in power series form). The measure of 4.18 then becomes equivalent to the L_2 norm, where

$$L_2^2 = \int_{-\pi}^{+\pi} |V(\theta)|^2 \frac{d\theta}{2\pi} \qquad 4.19$$

for small $V(\theta)$.

This in turn can be approximated by a truncated cepstral distance measure

$$U = \sum_{k=-p}^{P} (C_k - C'_k)^2 \qquad 4.20$$

where $\{C_k\}$ are the cepstral coefficients of $\sigma/A(Z)$ defined by

$$\ln[\sigma^2/|A(e^{j\theta})|^2] = \sum_{-\infty}^{\infty} C_k e^{-jk\theta} \qquad 4.21a$$

with

$$C_0 = \ln[\sigma^2] \qquad 4.21b$$

and

$$C_{-k} = C_k \qquad 4.21c$$

and $\{C'_k\}$ are likewise defined for $\sigma'/A'(Z)$.

Variants of 4.18b can be obtained by constraining σ and σ':

For $\sigma = \sigma'$ we obtain the likelihood ratio measure

$$L_R = \frac{\delta}{\alpha} -1 \qquad 4.22$$

If we set $\partial^{IS}/\partial(\sigma/\sigma') = 0$ we obtain the gain optimized, Itakura, or log likelihood ratio

$$LLR = \ln \left(\frac{\delta}{\alpha}\right) \qquad 4.23$$

Changing the roles of the coded model with the uncoded model one obtains

$$IS' = \int_{-\pi}^{\pi} [e^{-V(\theta)} + V(\theta)-1] \frac{d\theta}{2\pi} \qquad 4.24$$

And the average of IS and IS' is related to the cosh measure

$$C = \ln [1 + B + \sqrt{B(2+B)}] \qquad 4.25a$$

with

$$B = \frac{1}{2} [IS + IS'] \qquad 4.25b$$

which is symmetrical to $V(\theta)$ as opposed to either IS or IS'. Gain normalised and gain optimized versions of the cosh measure can also be derived. In a comparison study

[17] between subjective quality measurements and a selection of the above measures, the cepstral distance measure was found to correlate best with subjective results for PCM, ADM (adaptive delta modulation) and ADPCM coders although the correlation was not particularly good for ATC, APC, bit error and overload conditions.

4.2.2 Perceptually Motivated Measures

From the above discussion it should become clear that a vast range of distortion models has been proposed in the literature. The number of different models is a reflection upon their lack of robustness to different coding conditions. Some models are useful (in that they correlate well with subjective measurements) in one coding situation but fail to predict results when the coders involved or the speech material used is changed. For these new conditions a new appropriate measure can be found, but again, there is no guarantee that this measure will be robust to other coding conditions. One way to overcome this problem is to include in the test as many and diverse degredations in the speech signal as possible. Several measures can then be used to find the best one that mostly predicts subjective results on the totality of conditions. In addition, a composite of various objective measures (e.g. through regression analysis) can be found which accounts for the subjective results better than any of the individual measures [6, 7, 8]. Clearly, by using a model with a large enough number of parameters, any subjective result can be simulated. There is still no guarantee though that the model will perform well under conditions outside the "training data" i.e. the subjective results used for its derivation. Moreover, the number of free parameters required may be quite large for a wide range of coding conditions.

For all the above reasons it is beneficial to base a distortion model on knowledge about the auditory mechanism. In a sense, we are trying to model not the human response to a particular condition, but the actual mechanism that results in the auditory response given a particular stimulus. If this can be done successfully, then a distortion measure based on such a model will be able to predict and postdict any set of subjective results, since by design, the response of the model will be the same as that of the average human subject or observer. This is because what is modelled is the mechanism, not the response.

4.2.3 Modelling the Auditory Mechanism

The first step to model the auditory mechanism is to obtain the "Auditory Spectrum" or Excitation Pattern of a particular sound. As was shown before this involves a transformation of the frequency axis and a certain envelope smoothing with equal width frequency windows in the critical band domain. Two general approaches have been used to obtain an equivalent auditory spectrum from a signal. One is through the use of a discrete fourier transform (DFT) and the other through a filterbank. A DFT that is evaluated through short-time windows, and the output of a filterbank can give identical results if the impulse response of the filters corresponds to the window used in the DFT.

For the following discussion, for the DFT method, a fairly large time window will be assumed (20 msec) in order to obtain a fine frequency resolution power density spectrum on a critical band scale. The excitation pattern is then obtained by convolution of the power spectrum with a "spreading" function as in [13]. Because the convolution is performed on the power density and not on the complex spectrum, the result is that phase is only taken into account within the interval of the equivalent frequency window of the initial DFT (i.e. ~ 50 Hz at all frequencies). This results in a spectrum which is very smooth and free from pitch structured peaks and valleys.

The alternative is to use a filterbank. One or two filters per bark are adequate to describe the critical band spectrum. Since no analytical approach is available in the literature, a critical band filterbank was designed. As a first approximation to the critical band windows the following expression for the time window was used:

$$h(t) = \frac{(2\pi bt)^{n+1}}{tn!} \cdot e^{-2\pi bt} \qquad 4.26$$

its frequency response is given by

$$H(f) = \frac{1}{(1+j\frac{f}{b})^{n+1}} \qquad 4.27$$

and

$$10\log_{10}|H(f)|^2 = -10(n+1)\log_{10}[(\frac{f}{b})^2 + 1] \qquad 4.28$$

Note that this time window or filter is symmetric on a linear frequency scale.

The filters have two free parameters. By requiring that the three dB bandwidth is equal to the critical bandwidth D corresponding to its centre frequency we set

$$b = D/2 \cdot \sqrt{(\sqrt[n+1]{2} -1)} \qquad 4.29$$

and therefore reduce the number of free parameters to one, the value for n. This value is related to the slopes of the filters. To chose n, a visual fit was made to Zwicker's excitation pattern corresponding to 40 dB SPL (filter a) as shown in figure 2.4-2 in chapter 2. This value was chosen so that it would provide a reasonable fit to our moderately asymmetric (in the critical band - CB domain) amplitude response. The amplitude response is shown in figure 4.5 in dashed lines. Crosses indicate the 40 dB SPL excitation pattern from Zwicker whereas, in solid lines the spreading function as defined by Schroeder et al [13] is shown. A suitable value for n was found to be 5. The plot in figure 4.5 corresponds to this value. The excitation pattern corresponding to 80 dB SPL is shown in figure 4.6 (crosses). Schroeder's spreading function is again shown in solid. The dashed curve is our filter's amplitude response "peak detected" using the formula

$$B(k+1) = B(k) \text{ when } B(k+1) \geqslant B(k) \qquad 4.30a$$

and

$$B(k+1) = B(k) * EXP\ (-dx/C) \text{ when } B(k+1) < B(k) \qquad 4.30b$$

B is the amplitude spectrum coefficient at critical band rate k, dx the constant distance in critical bands between k and k+1 and C a constant determining the shallow slopes of the spectrum. The rather narrow filters resulting from our time windowing collect the sound pressure per critical band. Amplitude dependent slopes can then be fitted by changing C as a function of B(k), at the transition point, when B(k) = B(k+1) (as for filter b). The well known amplitude nonlinearities of the upper slope (accessory excitation) could then be modelled in this way.

The filters obtained from 4.26-4.29 (using n = 5) are relatively invariant in the critical band domain with respect to the centre frequency due to the normalization in 4.29. The resulting time windows though are very different for different centre frequencies since their width is inversely related to their frequency width, which in turn, is proportional to the critical bandwidth. The impulse responses of the filters (i.e. equation 4.26) for the windows corresponding to successive critical bands 1 to 17 are shown in figure 4.7. Having designed appropriate windows corresponding to critical bands, these can be used to obtain a critical band spectrum.

The spectrum is obtained by transforming the signal via a "Critical Band Transform" (CBT) matrix which operates on the time domain samples and produces the Excitation Spectrum without the need for a convolution in the CB domain. Convolution in the CB domain is wasteful since the short-time power spectrum of the speech signal has to be sampled to an unnecessarily high rate to produce an accurate convolution result which itself need only be represented by a smaller number of samples.

For a signal sampled in the time domain at a sampling period T the (continuous) power density spectrum can be represented by

$$S(f) = \left| \sum_{i=0}^{n-1} R(i)*EXP(-j2\Pi fiT) \right|^2 \qquad 4.31$$

valid between f=0 and f=1/2T. {R(i)} represent n successive sample values of the input signal and f is the frequency. Convolution in the CB (or frequency) domain can be substituted by multiplication by a suitable time window in the time domain. Because the width of the critical bands varies continuously with centre frequency, the time window is a function of the frequency at which the convolution result is calculated. The complex spectrum is then given by:

$$A(f) = \sum_{i=0}^{n-1} R(i)*W(i,f)*EXP(-j2\Pi fiT) \qquad 4.32$$

A discrete spectrum can be obtained by sampling the above continuous spectrum at any specified intervals. The frequency variable "f" can be expressed as a function of the corresponding critical band number X:

$$f = 650 * \sinh (X/7) \qquad 4.33a$$

and sampled at equal intervals of X i.e.

$$f = 650 * \sinh (kX/7) \qquad 4.33b$$

where X is a fixed constant.

The discrete spectrum is then given by

$$A(k) = \sum_{i=0}^{n-1} R(i)*W(i,k)*EXP(-j2\Pi f(kX)iT) \qquad 4.34a$$

where f(kX) means "f a function of kX" and not multiplication. In theory, the speech samples are time windowed prior to the transformation. In this case the time window coefficients can be incorporated into the transform, forming the new transform coefficients

$$W(i,k)*EXP(-j2\Pi f(kX)iT). \qquad 4.34b$$

The whole operation can then be presented in matrix form

$$A = WE * R \qquad 4.35$$

where WE is the transform matrix, R the input samples vector and A the critical band spectrum vector. The power spectrum coefficients are simply given by $S(K) = |A(k)|^2$. The whole operation is an exact equivalent to filtering by phase complementary band-pass filters centered at f(kX) (Flanagan pp145 [18]), and the form of 4.35 is similar to a DFT. The window W(i,k) is given by a sampled version of 4.26 such that 4.29 is satisfied at each frequency index k. The window appears in the matrix time reversed due to the convolution operation as shown in figure 4.7.

A fast implementation of 4.35 can be obtained by noting that the relevant windows W(i,k) decay to quite small values after a certain number of samples, especially for windows corresponding to high centre frequencies. The impulse response of 4.26 can therefore be truncated to an appropriate number of trems. A reasonable criterion is to truncate the series when the window value $h(t_T)$ is a small fraction of its maximum value $max(h(t_w))$. For a ratio of h(t)/max(h(t)) of around 300 the following number of samples need be retained:

Critical band	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Samples	159	159	147	137	126	115	104	93	82	72	64	56	48	42	37	32	28

Table 4.T2

for speech sampled at 8KHz. The maximum value nearly corresponds to 20 msec whereas for higher centre frequencies the time window is much shorter.

The resulting auditory spectrum is very different from the one obtained using the DFT approach. Recall that in the DFT approach [13], the power spectrum is convolved with the "spreading function". The filterbank spectrum gives an increased time resolution as a result of using impulse responses which are shorter than the time window in the DFT. It also gives an increased frequency resolution since the resulting spectrum is the convolution of the filters's response with the complex spectrum of the signal. This can clearly be seen by comparing figures 4.8 and 4.9.

In figure 4.8, column one shows the time waveform (20msec in each plot), the second column the power spectrum obtained via a 20 msec (warped) DFT, the third column, the convolved power spectrum (excitation) and the fourth the corresponding specific loudness pattern, obtained from the excitation through a corresponding power function with exponent 0.25. All the spectra extend from 0 to 17 bark, are linear on the bark scale and have a resolution of about 10 samples/bark. Each successive row is produced by sliding the time windows by 2 msec. It can be seen that via the DFT approach a smooth spectrum is obtained which changes little in the course of 8 msec. In figure 4.9 the four patterns shown represent the specific loudness patterns obtained from the filterbank and correspond to the same block of speech as the ones in the previous figure. It is clear that more information is preserved in the spectrum. In addition there is a lot of change in the spectra in the course of the 8 msec as the pitch harmonics, summed into the corresponding filters, fall in and out of phase. Here the thin curve corresponds to the filters with a slope as shown in figure 4.5 whilst the thick curve corresponds to the filters with added slopes as in figure 4.6. The spectrum obtained from the filterbank is not a true envelope of the fine spectrum but also reveals the fine structure at low frequencies. In addition, it can be seen that the whole spectrum is periodic, i.e. the spectra obtained from time windows separated by one period are similar. This is a new result, and to the knowledge of the author not mentioned anywhere in the literature. It could provide a clue as to how pitch information is coded by the auditory mechanism i.e. through a periodic pattern matching procedure of the whole spectrum. This is a viable alternative to either place or volley theories and, in fact, is a composite of both.

These patterns do not take into account any effects of temporal masking, other than those dictated by the width of the corresponding time windows. Such masking is expected to fill at least partially any valleys due to the preceding peaks at the same frequency location. Zwicker [19, 20] gives an analog circuit which he uses to simulate temporal masking. We implemented this nonlinear-low-pass device (NLD) through difference equations. The effect of including nonsimultaneous masking can be seen by comparing figures 4.10 with 4.11 and 4.12. Figure 4.10 shows a series of loudness patterns taken at 0.5 msec intervals without the NLD, using filter a (matched to 40 dB SPL) in figure 4.11 and filter b in figure 4.12 (matched to 80 dB SPL). It can be seen that some of the resolution is lost by using the NLD (nonsimultaneous masking) and filter b (equivalent to an increase in the upper accessory excitation) but the spectra are still different from the ones obtained via the DFT approach. The effect of the upper accessory excitation is relatively small. This is because the effect occurs at the skirts of the filters where the contribution to the spectrum is small. The periodic repetition of the spectra in time is also evident in figures 4.10-4.12

4.2.4 Separating the distortion from the signal

Having obtained equivalent auditory spectra it is necessary to investigate how to measure distortion from these auditory representations. A subjective listener does not use the original signal and the coded signal simultaneously. He listens to the coded signal and only infers from the original from his past experience (either immediate or long-term).

In a model of distortion, the noise has to be extracted from the coded signal using the original signal. This can be done either after the auditory spectra are obtained or before (i.e. from the time waveform. Schroeder et al [13] extract the noise from the time waveform. The equivalent of this operation in the frequency domain is to extract the noise spectrum from the encoded speech spectrum with "infinite" resolution in both amplitude and phase. This does not reflect the properties of the Peripheral Auditory System (PAS). It is not known how the brain compares the coded speech with the original but it can only draw a comparison on the (reduced) information passed by the (PAS). It would therefore be more reasonable to compare the specific loudness spectrum of the coded speech with the specific loudness spectrum of the original speech rather than extracting the noise prior to the processing. Therefore, in the model developed here, the distortion is extracted from the equivalent spectra of the coded and original signal.

4.2.5 Measuring the Distortion

Finally, after obtaining the two spectra, one corresponding to the original and the other to the coded signal, the amount of distortion has to be evaluated, taking into account the effects of masking. We will now go over some of the model elements that have already been presented in section 2.4 of chapter 2.

From Steven's law:

$$\frac{dN}{N} = k \frac{dI}{I} \qquad 4.36$$

where N is the loudness and I the intensity.

In the case of a wideband signal Zwicker proposed:

$$\frac{dN'}{N'} = k \frac{dE}{E} \qquad 4.37$$

where E is the excitation and N' the specific loudness (the value of loudness from a single critical band).

From the above differential equation we derive:

$$N' = (E)^{k} \qquad 4.38$$

in arbitrary units.

In the case of two different sounds being present

$$N_n' + N_o' = (E_n + E_o)^{k} \qquad 4.39$$

where n denotes the noise, 0 the original and c the coded signal. This can be written as:

$$N_n' = (E_c)^{k} - N_o' \qquad 4.40a$$

or

$$N_n' = N_c' - N_o' \qquad 4.40b$$

It is known that for N_n' to be zero it is not necessary that $N_c' = N_o'$ but only that

$$N_n' \leqslant N_{th}' \qquad 4.41$$

where N'_{th} is the value of N' from 4.38 for which the excitation E produces an inaudible effect.

Therefore:

$$N_n' = N_c'-N_o'-N_{th}' \qquad N_c'-N_o \geqslant N_{th}' \qquad 4.42a$$

$$N_n' = 0 \qquad N_c'-N_o'<N_{th}' \qquad 4.42b$$

N'_{th} is a function of the excitation E_o (or the specific loudness N_o').

In the case of noise masking a tone, the threshold N_{th}' is about 3 dB lower than N'_o (loudness in dB is given by $40\log_{10}N$, by analogy, from 4.38 with k = 0.25) but in the opposite situation a tone has to be some 25 dB higher than the noise to make the noise inaudible.

Using the very simple formula of 4.42 the effects of masking can be explained. First assume that N_{th}' can be taken to be zero for simplicity. With reference to figure 4.13, consider that $dE_1 = dE_2$ and that these represent the additional excitation due to noise which is added to E_1 and E_2. Note that E_1 and E_2 depend on the power of the signal that falls into the particular critical band (CB) in question, and therefore, the wider the filter, the higher E_1 and E_2. The effects of the amplitude nonlinearity (i.e. the exponent $K\tilde{}$ 0.25) is to make $dN'_2<dN'_1$ when $E_2 > E_1$ even though $dE_1 = dE_2$. If dE is a noise band in the centre of the CB and E comes from the speech harmonics that fall into the same CB, then E will mask dE even if the harmonics and the noise do not occupy exactly the same frequency location. The masking will be more effective the larger E is (provided dE_1, the maskee, is constant) either due to a wider critical bandwidth or larger power in the masker. The effect of N'_{th} is to reduce the loudness of the maskee even further or make it inaudible.

The problem of extracting the distortion from the auditory spectrum is a separate and to the author's view a more difficult problem than obtaining the equivalent auditory spectrum. This is mainly because there is no counterpart in the real system we try to simulate. The way different workers have chosen to model the process reflects the way each worker views the problem. Schroeder et al [13] for example tried to tie the problem to the traditional experiments on masking (using tones and noise bands) and view the problem as the masking of noise by the speech signal. In diriving their formulas, data from classical masking experiments were used, where there was a distinct masker (pulsed tone) and a distinct maskee (a noise band). The problem with this approach is that in most cases in speech compression, there is a high degree of correlation between the resulting noise and the signal and, quite often, a separate noise cannot be heard as the noise fuses with the signal to produce the distorted speech.

We prefer to view the problem as a pattern recognition exercise: when two coded speech signals are compared for distortion, a comparison would be made between each one and the original and the one which is the most dissimilar to the original is judged to be more distorted. Even when the original is not used for the comparison, human listeners are so overtrained in uncoded speech that such a comparison can somehow be made.

The stages beyond the peripheral auditory system are presented with a surface in a three dimensional space. The first dimension specifies the place on the Basilar Membrane (BM) at which a sound is mapped, the second is the neural "magnitude" that represents the disturbance and the third is the time course of the neural amplitude at the specified place on the BM.

Our basic assumption is that speech distortion is monotonically related to the dissimilarity between the two three-dimensional surfaces described by the above dimensions, one surface corresponding to the uncoded, original speech signal, the other to the coded signal. At this stage it is assumed that these differences carry equal weight irrespective of their position along the first dimension.

4.2.6 The complete models

After comparing the realization of different stages of the model using different signal processing techniques we now describe two algorithms, by Schroeder et al [13] and the one developed by the author, with the aid of two block diagrams.

The model by Schroeder et al is shown in figure 4.14. The various operations are as follows:

1. The noise is extracted from the coded signal
2. The speech signal S(t) is transformed into a warped power spectrum S(X) by means of a DFT and a power density correction factor df/dX.
3. S(X) is then convolved with B(X) to produce E(X), the excitation spectrum.
4. The excitation spectrum is then amplitude compressed to give the specific loudness N'_s, which, when integrated over the bark scale will give L_s, the loudness of the signal.
5. The above steps are followed for the noise signal N(t) with the difference that the excitation of the noise is reduced by the excitation of the signal.
6. Finally, the ratio of the two loudnesses gives a measure of the distortion in the signal.

The distortion measure designed by the author is given in figure 4.15. The original and coded signals are processed in parallel, over identical paths, up until the point where the equivalent auditory spectra are compared The operations are as follows (signal refers to both coded and original):

1. The signals are passed through a critical band filterbank employing phase complimentary bandpass filtering (PCBPF) to obtain the relatively smooth spectra |F(Z)| per critical band.

2. The slopes are adjusted on the high frequency side if necessary to produce the excitation patterns E(Z).
3. The excitation patterns are amplitude compressed to give N(Z).
4. The effects of "nonsimultaneous masking" can be included in this stage (i.e. the spilling over of the excitation into time regions where the signal is not actually present).
5. The two Auditory Spectra are compared.

4.2.7 Objective and Subjective Tests

4.2.7.1 Test material

In order to assess the performance of the two models, appropriately coded speech segments (files) were selected. For some of the speech files, the effects of auditory masking were exploited during the coding procedure. For these files the perceptual distortion was much smaller than conventional objective measures would indicate. Specifically, we compare two different ADPCM noise shaping schemes as a function of the noise factor, (equation 3.2-148). When the noise factor is equal to zero complete noise shaping is effected. When the noise factor is equal to one no noise shaping takes place and in this case the two schemes converge into one coding algorithm. This is an ADPCM backward block adaptive algorithm implemented at 16 kbits/sec. [21] Coded speech files under the first coding algorithm will be referenced as bm1p0, bm0p8, bm0p6 etc. the fraction denoting the noise factor (p represents the decimal point). The second algorithm will be referenced as 1m1p0, 1m0p7 etc. As mentioned before bm1p0 and 1m1p0 represent the same coded speech file and will both be referenced as bba1p0. Also 6 and 7 bits μ-law pcm coded speech was used in the comparison, denoted by PCM6 and PCM7 respectively. Two sentences were used in the comparison spoken by the same male person. These were treated as one continuous speech segment after silence deletion. The two sentences were "There was an old man called Michael Finnegan. He grew whiskers on his chinnagen".

4.2.7.2 Subjective Tests

In informal subjective tests that were carried out [21] the bm coding scheme was judged to be superior to the 1m coding scheme at the same noise factor. Further the 1m noise scheme gave the best performance at a noise factor around 0.6 whilst for the bm scheme improvement continues for smaller noise factors giving the best result at a noise factor of around 0.4 to 0.2. It should be noted that as the noise factor for the best result is approached the rate of improvement diminishes. The perceptual minimum distortion seems to represent a smooth minimum in the distortion versus noise factor curve. Comparisons with the PCM schemes revealed that 1m0p7 is judged to be equivalent to PCM7 whilst bm0p2 is judged to be superior to PCM7. The scheme that employs no noise shaping, bba1p0, is judged to be slightly worse than PCM6.

4.2.7.3 Objective Tests

Various algorithms were used to obtain objective results on the previous coded files. These are shown in figures 4.16 to 4.30. Since only the relative magnitudes of the distortions are of interest all distortion values are shown normalised by the distortion obtained for the bba1p0 scheme for each algorithm.

The SNRSEG performance is shown in figure 4.16. It shows that the bm scheme performs better than the 1m scheme which agrees with the subjective results. It also shows that there is a monotonic decrease in quality as the noise factor is reduced which is in disagreement with the subjective results. It also shows that even PCM6 is better than any of the ADPCM files which is in gross disagreement with the subjective results.

The performance of Schroeder's model is shown in figure 4.17. It correctly shows that there is no monotonic relationship between the distortion and the noise factor but that there is a minimum distortion between the extreme values of the noise factor. Note though that the relationship between the ADPCM schemes and the PCM schemes is shown to be essentially the same as that from the SNRSEG.

In our distortion measure the distortion for each time slice is given by:

$$d(t) = \sum_{z=1}^{n} F[SL_o(z,t),\ SL_c(z,t)] \qquad 4.43$$

and the accumulated distance by:

$$D = \sum_{t=1}^{T} d(t) \qquad 4.44$$

Here F is a function of the specific loudness of the original (SL_o) and the coded (SL_c) at time t and critical band position z. No sliding was used in the time windows i.e. the time interval between successive time frames is 20 msec. The filter used was filter b (matched to 80 db SPL). We drop the index t (for time) although the function F is still defined for a specific time instant.

The function used to obtain figure 4.18 was a straight difference function:

$$F = [SL_c(z) - SL_o(z)] \qquad 4.45$$

It can be seen that some improvement in agreement has been achieved. The files (speech segments) with middle values of distortion are shown to have less distortion than PCM6.

The results shown in figure 4.19 are similar to the ones in 4.18 apart form that the files with smaller noise factors show a slightly increased distortion. The formula used here was:

$$F = [\frac{SL_c(z)}{L_c} - \frac{SL_o(z)}{L_o}] \qquad 4.46$$

where L_c and L_o represent the loudnesses of the coded and original respectively.

An even better agreement with subjective results is given in figure 4.20. For these, the function used was

$$F = [SL_c(Z) - SL_o(z)] - SL_{th}(z) \qquad 4.47$$

were $SL_{th}(z)$ represents the threshold value as a function of the critical band index z. F is replaced by zero in the case where $[SL_c(z) - SL_o(z)] < SL_{th}(z)$. We evaluate $SL_{th}(z)$as:

$$SL_{th}(z) = TH*SL_o(z) \qquad 4.48$$

where

$$TH = 10^{(-32.0+0.75z)/40.0} \qquad 4.49$$

This value for the threshold was found to give results which were close to the subjective ones.

The same function F was used to obtain figure 4.21 but now the specific loudnesses are normalised by the corresponding loudness values as in 4.46. Again, the normalisation shows increased distortion in the files with low noise shaping factor (increased noise shaping).

The threshold can be considered as a sensitivity function: The higher the threshold, the less noise passes through at that particular frequency.

When the loudness of a sound is to be calculated the last stage of the calculation is a 'leaky' integration with a time constant of around 160 msec. The next three cases introduce such a low pass operation on the specific loudness before forming the function F. In figure 4.22 F was evaluated with no threshold and no normalisation. It can be seen that the results are similar to those in figure 4.20. With the normalisation introduced the results are shown in figure 4.23. Now, the normalisation reduces the distortion in the files with increased noise shaping. Finally, figure 4.24 shows the combined effects of the normalisation and threshold with the same threshold value as before. The distortion in the files with low noise factor is shown to be much lower than the one obtained from subjective results.

All the above measures used a function of F which was based on the difference of the specific loudnesses. We next consider three different forms for F. These can be considered as self normalising (i.e. a division by the instantaneous Loudnesses would not alter the result).

In the first one, F is a ratio measure:

$$F = \left[\frac{SL_c(z) - SL_o(z)}{SL_o(z)} \right] \qquad 4.50$$

This algorithm, shown in figure 4.25 underestimates the distortion introduced due to the increased noise shaping.

It over-emphasises the effects of masking and in this sense is the opposite of the SNRSEG measure.

Next, in figure 4.26 the results shown were obtained using a log ratio measure

$$F = |\ \log(\ SL_c(z)\ /\ SL_o(z)\ )\ | \qquad 4.51a$$

or

$$F = |\ \log(\ SL_c(z)\ ) - \log(\ SL_o(z)\ )| \qquad 4.51b$$

note that this form is exponent invariant i.e. it would give the same results whether specific loudnesses were used, power densities or spectral amplitudes.

The results are very similar to the ratio measure. This should be expected when $SL_c \sim SL_o$: Let $A=\log(SL_o)$ then,

$$\frac{dA}{dSL_o} = \frac{1}{SL_o} \quad \text{and} \quad dA = \frac{dSL_o}{SL_o} \qquad 4.52$$

now if $dA = |\log(SL_c) - \log(SL_o)|$ i.e. $dA = F$ then

$$F = \left|\ \frac{SL_c(z) - SL_o(z)}{SL_o(z)}\ \right| \qquad 4.53$$

i.e. the same expression as the ratio measure, provided $SL_c - SL_o << SL_o$.

Another possible form for F could be a variance measure:

$$F = \left|\ \frac{SL_c(z)}{SL_o(z)} - A\ \right| \qquad 4.54$$

where

$$A = \frac{1}{n} \sum_{z=1}^{n} \frac{SL_c(z)}{SL_o(z)} \qquad 4.55$$

Therefore if the spectra are multiples of each other the distortion is zero and the measure gives the variance of the ratio over the spectrum. The results are similar to the two previous ones considered above, figure 4.27. (Note that for A ~ 1 then again F = $|SL_c(z) - SL_o(z)|/SL_o(z)$.

The variance measure now explains why the above three algorithms underestimate the distortion in the files with the highly shaped noise: In this case the envelope of the coded signal will have been raised uniformly by the noise, making the envelope of the coded signal, on average, a multiple of the envelope of the original signal.

Next we investigated the effect of the amplitude compression in obtaining the specific loudness from the excitation spectrum. The function F is now:

$$F = \frac{SI_c(z)}{I_c} - \frac{SI_o(z)}{I_o} \quad , \quad SI = (SL)^4 \qquad 4.56$$

Therefore we replace each value of Specific Loudness by the corresponding value of Specific Intensity. Our measure now gives the power spectrum difference. The results of this algorithm are shown in figure 4.28. Note the similarity between this and the SNRSEG measure (taking into account the fact that the SNRSEG is shown onto logarithmic scales). This is of course a reminder of Parceval's Theorem although the equivalent time windows in the two measures are different.

Finally, to conclude this section we investigate the effect of using equal bandwidth filters instead of CB filters. As our bandwidth, we used the average value of all the CB's over the range of 0-3.4 KHz. This is about 200 Hz. The expression for F used was

$$F = |SL_c(z) - SL_o(z)| \quad \text{in figure 4.29} \qquad 4.57$$

$$F = \left| \frac{SL_c(z)}{L_c} - \frac{SL_o(z)}{L_o} \right| \quad \text{in figure 4.30} \qquad 4.58$$

The results are not markedly different form the ones obtained from the warped frequency cycle. Over some of the files, there is an improved performance compared to the warped case. The equal bandwidth case deserves special attention since the operation can be implemented using FFT's of a small analysis window (5~ 10 msec or 40 ~ 80 point) which would increase the execution speed of the algorithm. The discrepancies between the results obtained from this method and the ones from the warped scale could be corrected by using a modified threshold curve.

Concluding, we see that the most important operation in the algorithm is the amplitude compression. The use of a special filterbank seems of less importance as long as a suitable bandwidth is used for the analysis. (representing some kind of average of the values for the critical bandwidth over the frequency range concerned). Difference algorithms seem to perform better than ratio algorithms. These results should be considered as indicative rather than conclusive due to the small sample size (although it was carefully chosen). A much wider selection of distortions would have been required to express a definite preference of one algorithm over another or optimize any of the algorithm parameters such as time constants and thresholds. It would then be possible to compare the results to the outcome of formal subjective tests and then select that algorithm (or combination of) with the best correlation with the subjective tests. This would require the results of the subjective tests to be expressed on a number scale. This does not normally pose any problems with formal tests. Due to the unavailability of such subjective results we direct our attention into other aspects of our model.

4.2.8 Multidimensional representation

Due to the small number of files available it is difficult to derive any general results about our different algorithms. In order to try to get as much information out as possible we propose a multidimensional representation of the speech distortions. We not only form a distance measure between each of the distorted files and the original but we also form a distance measure between every possible pair of distorted files. Note that this is a direct follow up of our decision to view the process as a pattern recognition exercise which enables us to compare two distorted files together. The alternative approach which extracts the noise from the signal prior to the processing is no longer applicable since no basis is provided to compare two coded files together.

Any of the various distance algorithms which have been presented before can be used to provide the accumulated distances which represent the distortion measure: e.g. if SL(z,t,i) represents the z coefficient of the spectrum at time t of file i then a suitable distance between the same time-slice for different files is given by:

$$d(t,i,j)=\mathrm{sqrt}\{\sum_{1}^{n} [SL(z,t,i)-SL(z,t,j)]^2\} \qquad z=1,\ldots.n \qquad 4.59$$

and the accumulated distance by:

$$D(i,j) = \sum_{1}^{m} d(t,i,j) \qquad t=1,\ldots.m \qquad 4.60$$

Although this may be an appropriate measure for the 'audible' noise, perhaps a more suitable measure for distortion would be:

$$nd(t,i,j)=\mathrm{sqrt}(\sum_{1}^{n} [\frac{SL(z,t,i)}{L(t,i)} - \frac{SL(z,t,j)}{L(t,j)}]^{2}) \qquad z=1,\ldots.n \qquad 4.61$$

where

$$L(t,i) = \sum_{1}^{n} SL(z,t,i) \qquad z=1,\ldots.n \qquad 4.62$$

and

$$ND(i,j) = \sum_{1}^{n} nd(t,i,j) \qquad t=1,\ldots..m \qquad 4.63$$

Before, one of the files above would have been the original, but any two files can be compared and therefore the dissimilarity between different coded files can now be assessed.

The results from applying any of the above distortion measures to every possible pair of files is n dissimilarity (distance) indices for m files where n=m(m-1)/2. For the files we consider above this gives 91 indices for 14 files. Treating each file as a point in a multidimensional Eucledian space and each index as the corresponding Eucledian distance between them, we can represent those points in an (m-1)-dimensional space. This would reproduce the original distances exactly and, in our case, in a 13-dimensional space. Such a representation is not possible to be visualised and therefore we try to map this space into one of a reduced dimensionality whilst trying to alter the original Eucledian distances as little as possible. One statistical technique, namely Principal Components Analysis [22. 23. 24] operates as follows in order to achieve the above result:

First from the matrix of the interpoint distances it derives a matrix of coordinates which will exactly reproduce the original distances. Then from these, principal components are derived. The first Principal Component is that weighted combination of the original variables which maximizes the variance amongst the different files subject to the constraint that the sum of the squares of the weights are equal to unity.The nth Principal component is that combination of the original variables which maximizes the remaining variance subject to the same constraint as the first and in addition its scores on (the coordinates of) the speech files are uncorrelated with the previous n-1 components. These constraints make the transformation an orthogonal one which preserves Eucledian lengths. The resultant components account for progressively less in the variance of the data and therefore a—large number of the components can be discarded with little loss of accuracy. We formed three dimensional spaces for the above files.

Next we will present some results obtained using the above technique. A selection of algorithms from the ones presented in the previous section were used on the same original and coded files. An example is shown in figures 4.31a and 4.31b. These represent the dissimilarities between the 14 files already mentioned in section 4.2.7.1. Figure 4.31a shows the projections of the files onto the x-y plane whilst figure 4.31b shows the projections of the files onto the y-z plane. A translation of the origin can be performed without any effect on the Eucledian Distances. An obvious choice for the origin is the original signal. We treat the PCM files as a 'reference' distortion. As can be seen form the figures both the PCM files lie on a line originating from the origin. This signifies that they both have a similar type of distortion of different magnitude. Call this line the PCM axis. The similarity of the distortions between any of the files and the PCM type distortion is given by the angle its radious vector (with the origin at the original) makes with the PCM axis, whilst the overall distortion is given by the length of this radious vector. The x-y plane largely discriminates between the different ADPCM files: When the noise factor is close to 1.0 (i.e. no noise shaping) the distortion has a large component on the PCM axis which diminishes as the noise factor is reduced. As the noise factor diminishes another type of distortion becomes larger. This is related to roughness which appears in the noise shaped speech. Note that the ADPCM files are ordered automatically according to the noise shaping factor. This is much more pronounced for the 1m scheme which agrees with the subjective results. The z axis discriminates between the ADPCM and the PCM files as a whole. For the two figures just mentioned the algorithm used was a straight difference between the SLs with no threshold included and no normalisation. (i.e. equation 4.45).

The distances from any of the algorithms used are only significant to within a multiplicative constant. In order to compare the results from the different algorithms these constants (as well as the orientation of the axes) have to be matched together. This can be done by a method called Procrustes Rotation (PR) [23]. Assume that the rows of a matrix X give the coordinates of points P_i (i=1,2....n) produced by one algorithm and the rows of a matrix Y give the coordinates of the points Q_i (i=1,2,....n) produced by another algorithm, in such a way that corresponding rows refer to the same speech files. In this method the configuration Y is translated, orthogonally rotated and isotropically stretched until

$$M^2 = \text{sum } (P_i Q_i)^2 \qquad 4.64$$

is minimised. (Note that the above operations preserve the ratios of the original distances i.e. the distances to an arbitrary multiplicative constant). The centroid of both X and Y now coincides. The sum of squares amongst the X coordinates equals the sum of squares amongst the Y coordinates plus the residual sum of squares or, algebraically:

$$\text{Trace } (XX') = \text{Trace } (YY') + M_{pq}^{2} \qquad 4.65$$

We normalize the Trace(XX') to have a value of 1.0. Then, the magnitude of the residual gives an indication of the similarity of the two configurations. We have compared the configurations of the ratio algorithm with the log ratio algorithm (see the previous section). Treating the ratio algorithm as the fixed configuration, then, after PR on the log ratio logarithm results, the trace is 0.9986 and the residual 0.0014, which indicates that the two algorithms produce very similar configurations. Further, we compared the results with and without normalisation for the difference algorithms. The residual was now 0.0491 which is over an order of magnitude larger than the previous one, but still small. To conclude, we compared several algorithms to the same fixed algorithm. The fixed algorithm was the difference one, with a threshold value of -32.0+0.75z (no normalisation). The residual values for the different configurations are shown in Table 4.T3. Some of the configurations can be seen in figures (4.32 to 4.35). The above configuration was chosen as the fixed one since it performed closely to the subjective results. These and the previous multidimensional results were produced using the Genstat high level language [23].

It can be seen that the configuration with the specific intensities gives the largest residual, followed by the log-ratio and the smoothing in the time domain. The equal bandwidth algorithm performs relatively closely to the fixed configuration showing that the effects of the warped frequency scale are not of paramount importance.

4.2.9 Applications to coding - selection of threshold parameters

Using the lm scheme we placed our distortion measure in a feedback loop to automatically select a noise factor for each speech block. The algorithm had to select the one value from the four (0.2, 0.4, 0.6, 0.8) which would give the smallest distortion. Then, using the 'best' blocks obtained for the various filter memories the algorithm proceeds to the next block and again evaluates the distortion for the new set of choices. In the case of a tie, the program makes a random choice amongst the files with the same distortion value. Using a threshold value of -32.0+0.75z and a sliding interval of 5 msec, we obtained a speech quality which was slightly better than with any of the fixed noise shaping factors. The distribution of the chosen values versus the block number (each block is equivalent to 20 msec) is shown in figure 4.36. It can be seen that the algorithm favours the low noise shaping factors.

The same coding structure can be used to assess the success of any of the distortion measures mentioned above. In particular we used the lm scheme with the distortion measure of equations 4.47-4.49 in the feedback loop in order to determine the best values for parameters a and b in the threshold function in 4.49 given by

$$TH = 10^{(-a+bz)/40.0} \qquad 4.66$$

The best values were determined by exploring the whole range of values for a and b, and using successively narrower ranges around the most promising values. The best values were those that produced a speech quality that was judged to be better than the other combinations of values. A coded file was produced for each set of {a,b} and the best was determined through the iterative procedure above, from informal subjective tests (by the author). Some of the functions (-a+bz) are shown in figures 4.37 a-d. In all, over 50 different combinations were assessed. The value of -32.0+0.75z was found to give a performance close to optimum, in line with previous results.

4.3 Subjective measurements of coder Quality

Subjective quality measures have to meet a multitude of often conflicting criteria. An appropriate measure has to be valid, representative, and reliable in order to provide a useful index of suitability for a particular coder: The result should contain as much information as possible about the coder and as little as possible about the speech materials used, the subjects or the particular experimental technique that was employed. The above criteria are usually conflicting with the need to provide a convenient, easy and inexpensive to use measure.

A compromise can be found in the mean opinion score measure (MOS). In this procedure subjects evaluate the quality of short sentences on a five point category scale: bad, poor fair, good, excellent [25, 26, 27].

To minimize variability and to aid comparison with similar tests, reference distortions are included such as log-PCM distortions or modulated noise reference unit (MNRU) distortions. These latter distortions are produced by generating additive noise that is proportional to the instantaneous signal amplitude and then adding it to the speech signal. By varying the constant of proportionality, different amounts of distortion can be produced.

A more thorough and informative test is the Diagnostic Acceptability measure (DAM) [28]. In this test subjects are asked to rate a particular coding condition not only in terms of its overall quality or acceptability but also along scales of specific degredations. Examples of such degredations in terms of resulting speech impairment are: fluttering, (interrupted or amplitude modulated speech), thin (high pass speech), rasping (peak clipped speech), muffled (low pass speech), hissing, buzzing, babbling, rumbling. The scores of the subjects along such scales can be transformed to an overall acceptability measure, or into an acceptability measure of any of the particular degredations through appropriate transformations. The form of the transformations needed have been determined through a training phase. For the training phase, over 20 different coding conditions were used, over several sentences, resulting in about 18 hours (!) or distorted speech. Appropriate objective measures suited to measure each particular degredation can then be devised and a composite created which would agree with the overall acceptabiity ratings [29, 30, 31].

Another method to obtain ratings on different types of quality degredation is through Multidimensional Scaling (MDS) Procedures [8, 32-36]. These procedures are not unlike the Principal Components Analysis (PCA) encountered earlier in that they accept as input some form of "distances" between stimuli (here, distorted speech files) and, at the output they place the stimuli into a multidimensional space such that the original (input) "distances" are "preserved" or reproduced. As a byproduct, the psychophysical scales or dimensions used by the subject are obtained. The output therefore contains similar information to the DAM measure although the various degredation scales are themselves part of the algorithm's output and are not predetermined for the experiment as in the DAM. As opposed to PCA, the input to (Nonmetric) Multidimensional Scaling is not (Eucledian) distances but similarities or dissimilarities between the distorted files. Therefore the subjects can be asked to rate directly how similar (or dissimilar) two distortions are perceived on some scale or, alternatively, they can be presented with triads of distortions, say A, B and C. The subject is then asked to quote the two most similar distortions and the two most dissimilar distortions from each triad.

The solution is usually obtained by minimizing a predetermined criterion such as

$$\text{stress} = \left[\frac{\sum_j \sum_k (d_{jk} - \hat{d}_{jk})^2}{\sum_j \sum_k d_{jk}^{\,2}} \right]^{1/2} \qquad 4.67$$

j and k refer to two different coding situations, d_{jk} are the distances in the (metric) space (the "solutions") and d_{jk} are related to the input similarities by a nonincreasing monotonic function.

Stress is then minimized over all m-dimensional spaces (m is determined by choice) and over all nonincreasing monotonic functions.

In order to investigate the subjective multidimensionality of distortion, already established for the objective results through PCA, the following test was constructed from a selection of files used before (lm 0p0, lm0p2, lm0p4, lm0p6, lm0p8, bbalp0, PCM6, PCM7, Original). All possible trials were formed. The order of the presentation of the files was randomized within each triad and the order of presentation of the triads themselves was also randomized. This resulted in 84 triads. Next, three subjects (MA, NG, SS) were asked to record the most similar pair and the most dissimilar pair within each triad. The results were analysed using two different algorithms, MINISSA and TRISOSCAL [37]. Minissa is a general MDS program as described above. Trisoscal is similar to Minissa but additionally attempts to remove any anomalies introduced by the method used in collecting the results and is specific to triadic comparisons. The two solutions are shown in fig. 4.38 and 4.39. The two solutions are fairly similar.

Several similarities can also be noted between these results and the objective configurations obtained in a previous section (section 4.2.8):

First the ADPCM files are ordered according to the noise shaping factor. Note that the subjects had no prior knowledge of the coded schemes used.

Second, the PCM files and the original lie on the same line.

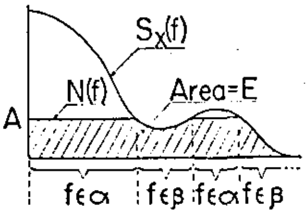
Third, the files with little noise shaping lie near the PCM axis.

The main difference between these and the objective results is that as the noise factor takes its extreme values the two ends of the plot curve towards each other to form a "bowl" shape instead of the right angle shape obtained in the objective results. Also note the distance of the original from the rest of the files. Although the differences could arise in the method used to collect and process the results, it seems that the first dimension in the subjective results is related to the overall distortion whereas the second dimension discriminates amongst the types of distortion.

Table 4.T3

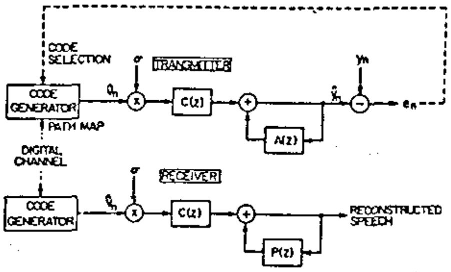
Trace and residuals of the different configurations

Algorithm	Trace, Residual	Figure
difference algorithm Thresh = -32+0.75z no normalization	1.0000, 0.0000	figures 4.32a and 4.32b
difference algorithm no threshold no normalization	0.9208, 0.0792	
specif. Power Densit. no threshold normalized	0.4436, 0.5564	
Equal Bandwidths no threshold no normalization	0.9228, 0.0772	figures 4.33a and 4.33b
Difference algorithm Thresh = 32+0.75z normalized	0.9300,-0.0700	figures 4.34a and 4.34b
160 msec on Spec. L. no threshold no normalization	0.8793, 0.1207	
Log ratio no threshold	0.8220, 0.1780	figures 4.35a and 4.35b



The relationship of noise and signal spectra.

Figure 4.1 [4]

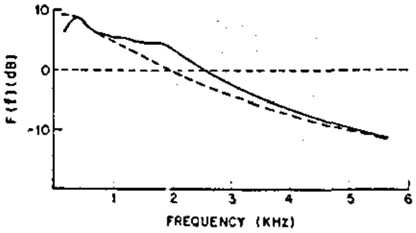


Block diagram of a coder for a spectrally modified speech signal with white quantizing noise.

Figure 4.2 [5]

Waveform Coder

Figure 4.3 [6]



Frequency weighting associated with articulation band and critical band warping.

Figure 4.4 [6]

FREQUENCY BANDS OF EQUAL CONTRIBUTION TO THE ARTICULATION INDEX (ALL ENTIRES IN Hz)

No.	*Limits*	*Mean*	*No.*	*Limits*	*Mean*
1.	200 to 330	270	11	1660 to 1830	1740
2.	330 to 430	380	12.	1830 to 2020	1920
3.	430 to 560	490	13.	2020 to 2240	2130
4.	560 to 700	630	14.	2240 to 2500	2370
5.	700 to 840	770	15.	2500 to 2820	2660
6.	840 to 1000	920	16.	2820 to 3200	3000
7.	1000 to 1150	1070	17.	3200 to 3650	3400
8	1150 to 1310	1230	18	3650 to 4250	3950
9.	1310 to 1480	1400	19	4250 to 5050	4650
10	1480 to 1660	1570	20.	5050 to 6100	5600

Table 4.T1 [6]

Figure 4.5

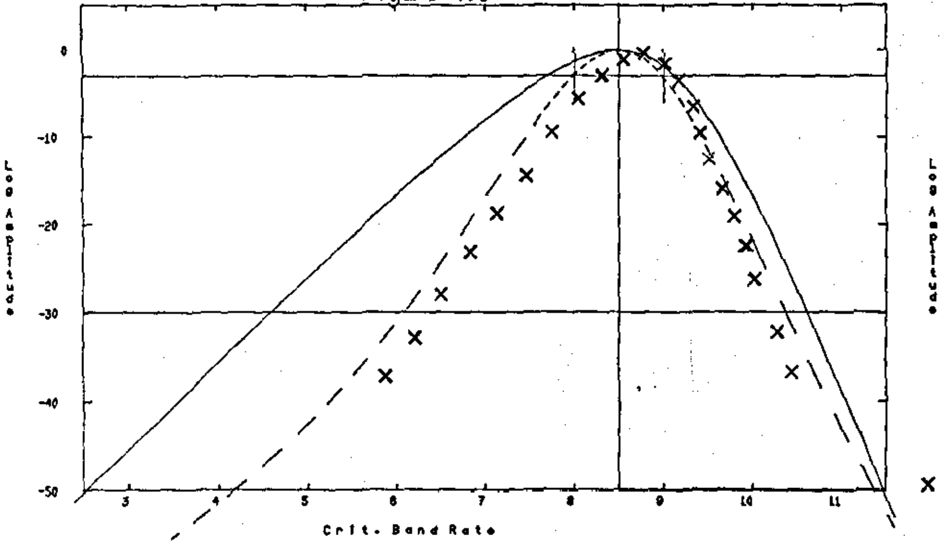


Figure 4.6

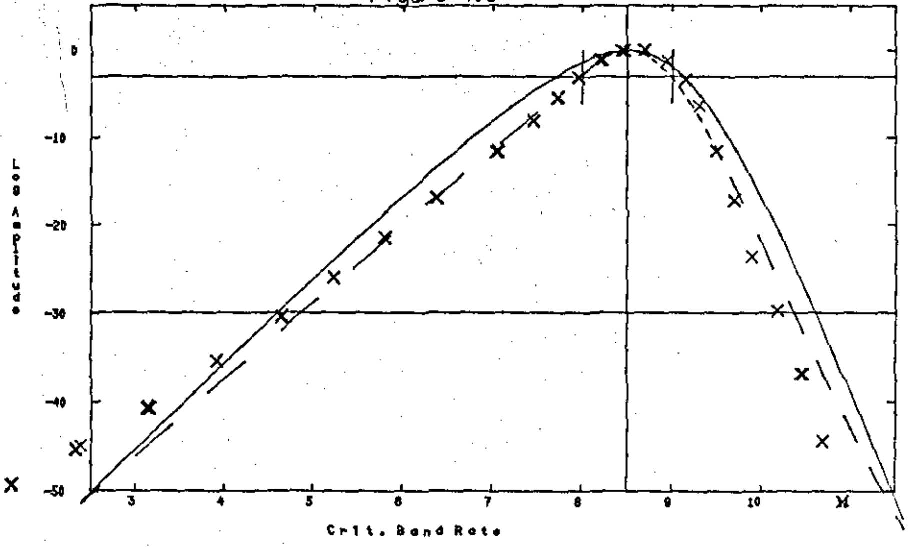


Figure 4.7

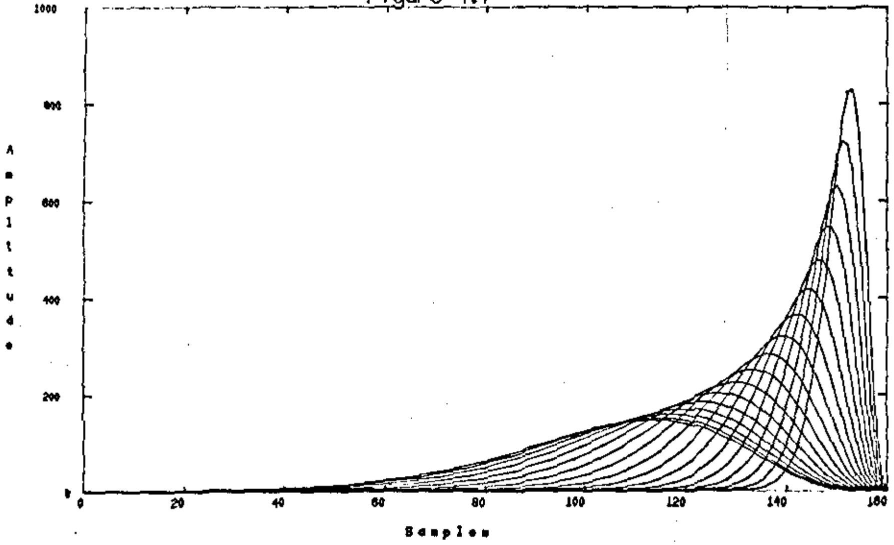
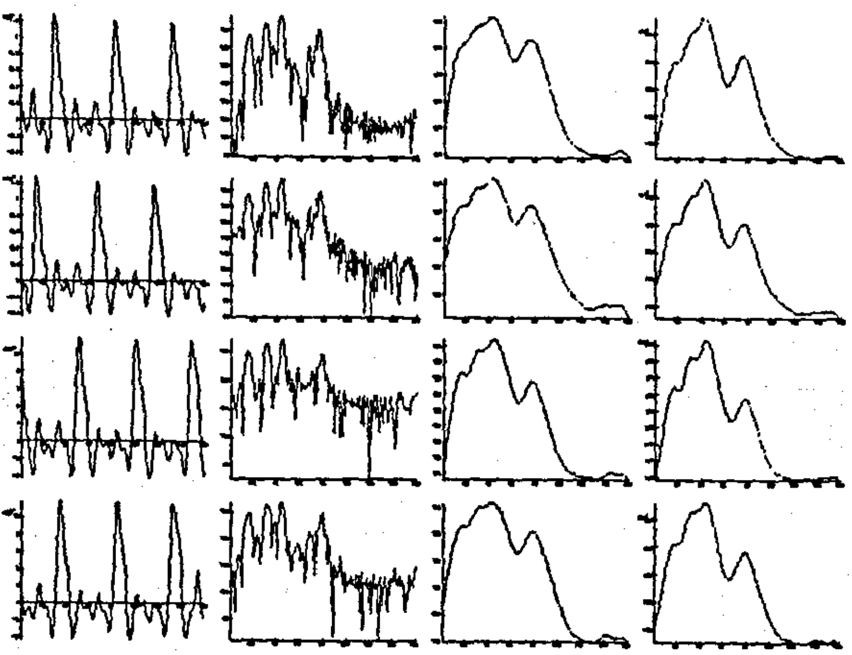


Figure 4.5 Filter amplitude responce as a function of frequency.
Solid line : Filter shape proposed by Schroeder et al.
Dashed line : Symmetric filter proposed by the author.
Crosses : Data from Zwicker. Curve at 40 dB SPL.

Figure 4.6 Filter amplitude responce as a function of frequency.
Solid line : Filter shape proposed by Schroeder et al.
Dashed line : Symmetric filter proposed by the author.
Crosses : Data from Zwicker. Curve at 80 dB SPL.

Figure 4.7 Impulse responce of the filter for the time windows corresponding to successive critical bands from 1 to 17.



time waveform   amplitude spectrum   excitation pattern   loudness pattern

Figure 4.8 The DFT approach

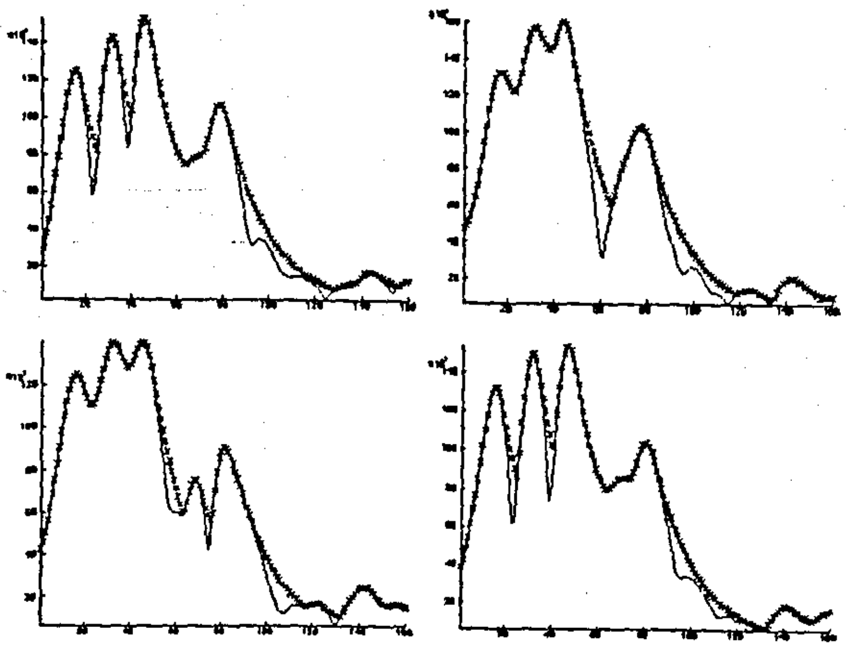


Figure 4.9 Successive specific loudness patterns.
Thin line : filters matched to 40 dB SPL.
Thick line : filters matched to 80 db SPL.

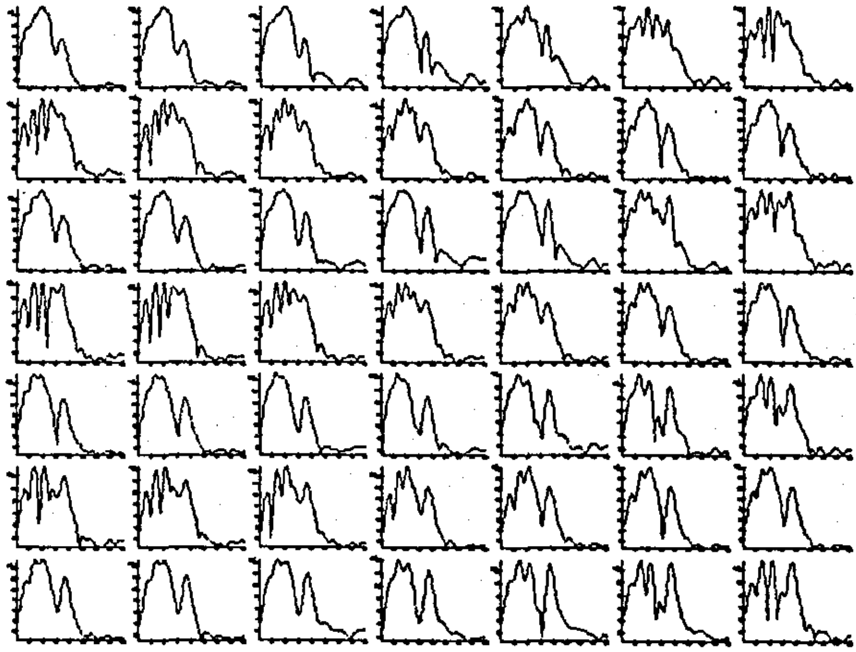


Figure 4.10 Specific Loudness patterns at 0.5 msec intervals obtained without using the NLD (nonlinear device–see text). Filter (a) (matched to 40 dB SPL) was used.

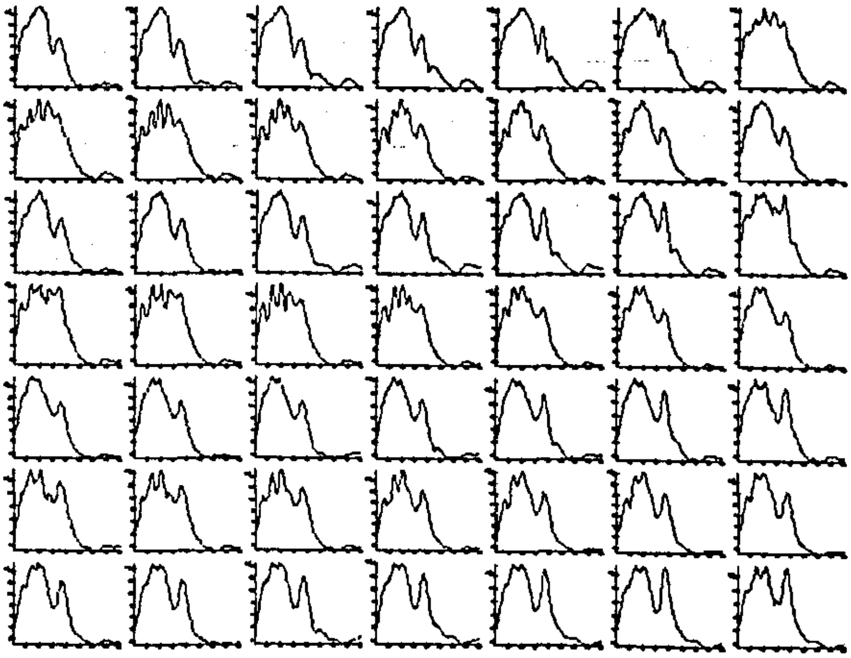


Figure 4.11 Specific Loudness patterns at 0.5 msec intervals obtained using the NLD (nonlinear device–see text). Filter (a) (matched to 40 dB SPL) was used.

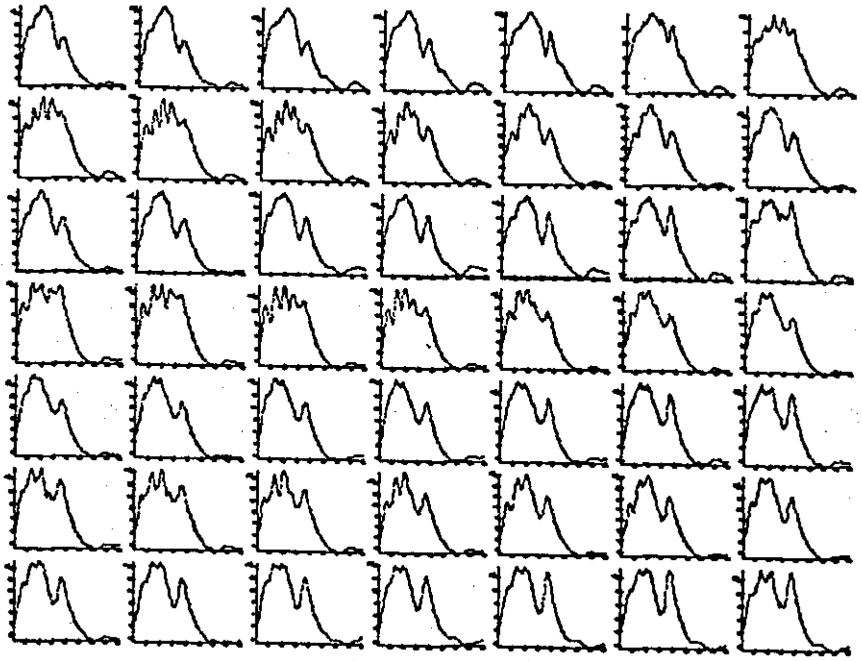


Figure 4.12 Specific Loudness patterns at 0.5 msec intervals obtained using the NLD (nonlinear device-see text). Filter (b) (matched to 80 dB SPL) was used.

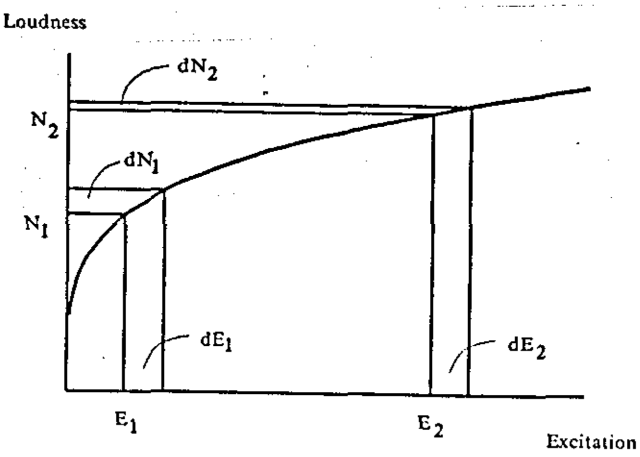


Figure 4.13 The amplitude compressing nonlinearity.

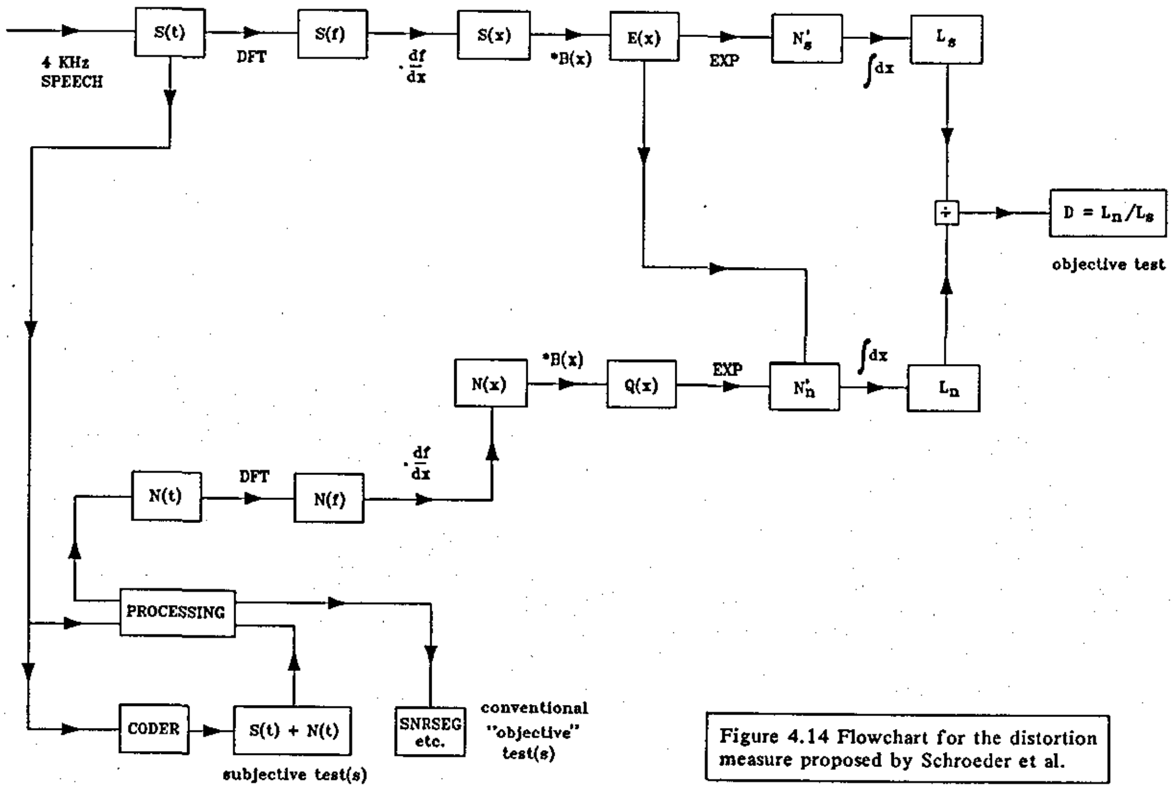


Figure 4.14 Flowchart for the distortion measure proposed by Schroeder et al.

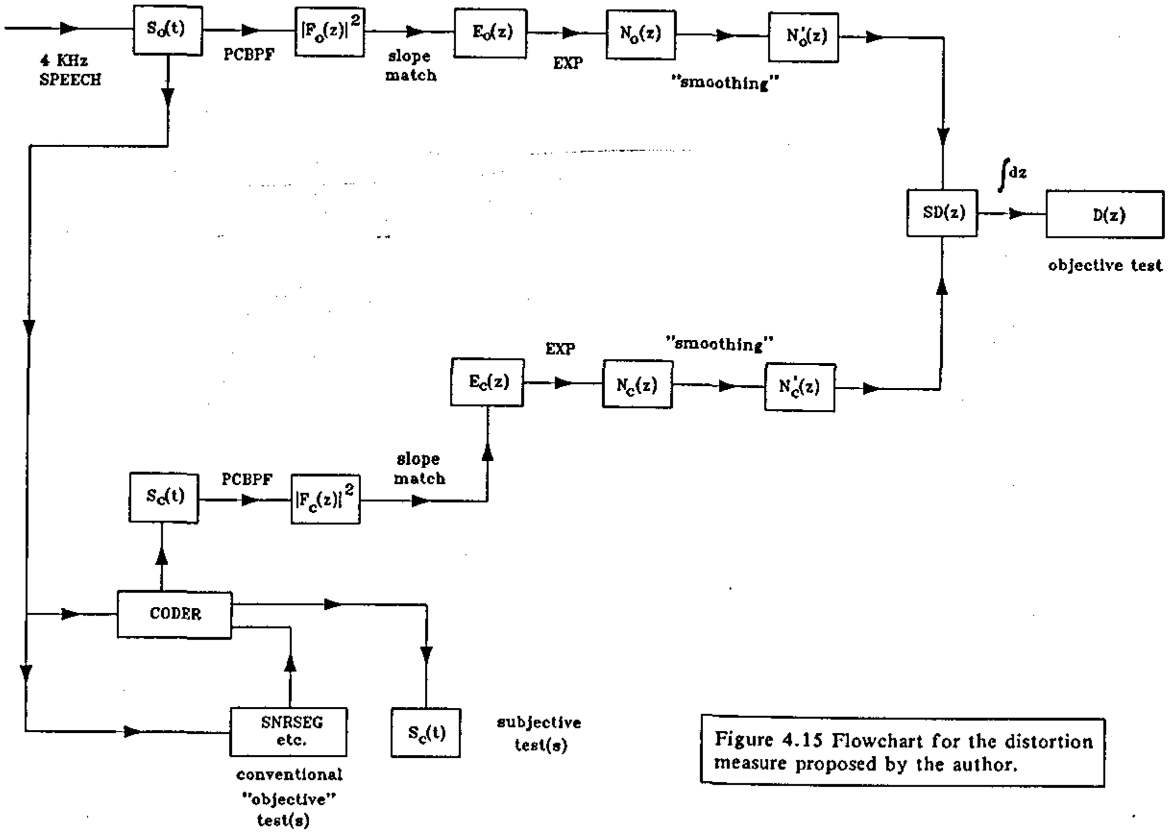


Figure 4.15 Flowchart for the distortion measure proposed by the author.

Figure 4.16 SNRSEG Measure.

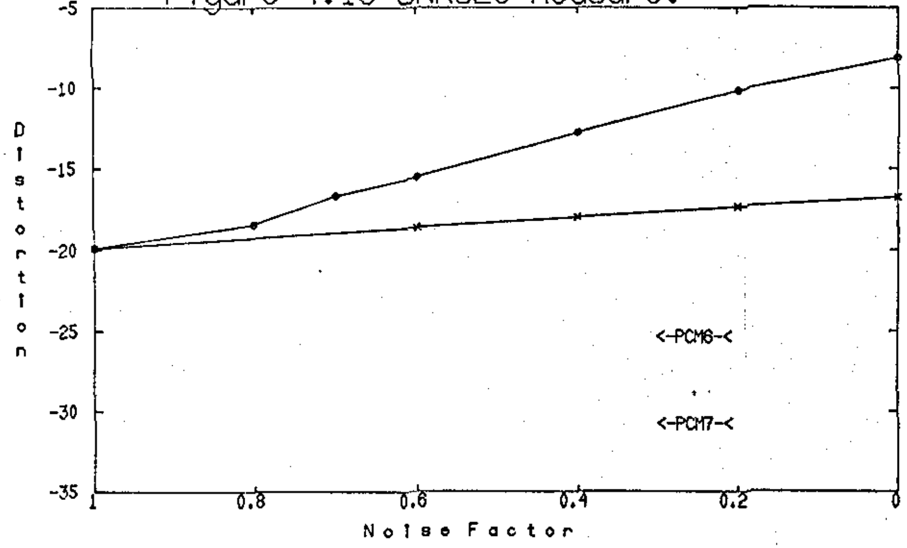


Figure 4.17 Schroeder et al measure.

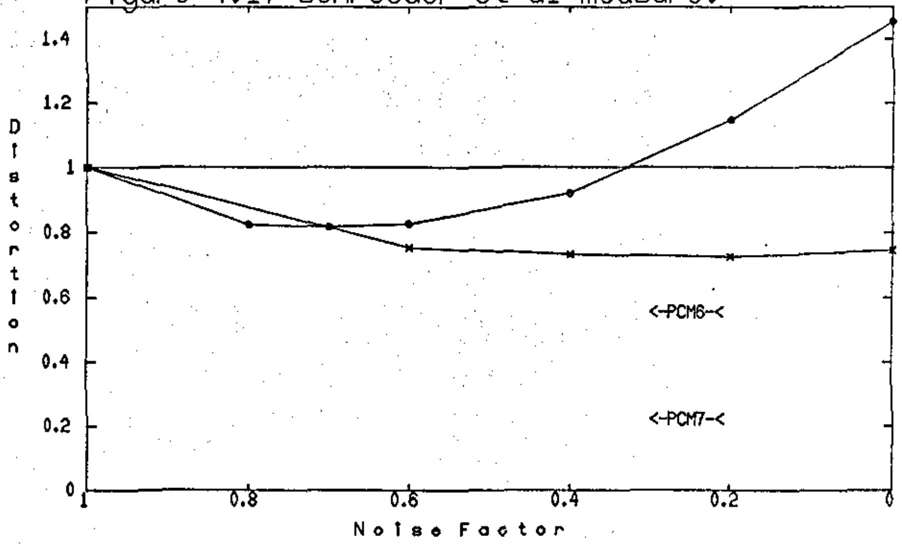


Figure 4.18 Difference measure, no threshold, no normalization.

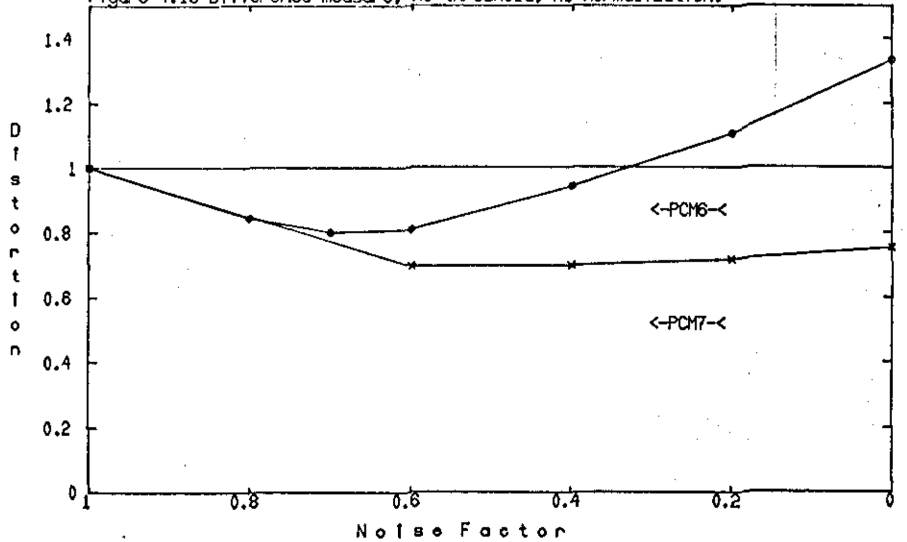


Figure 4.19 No threshold, normalized specific loudnesses.

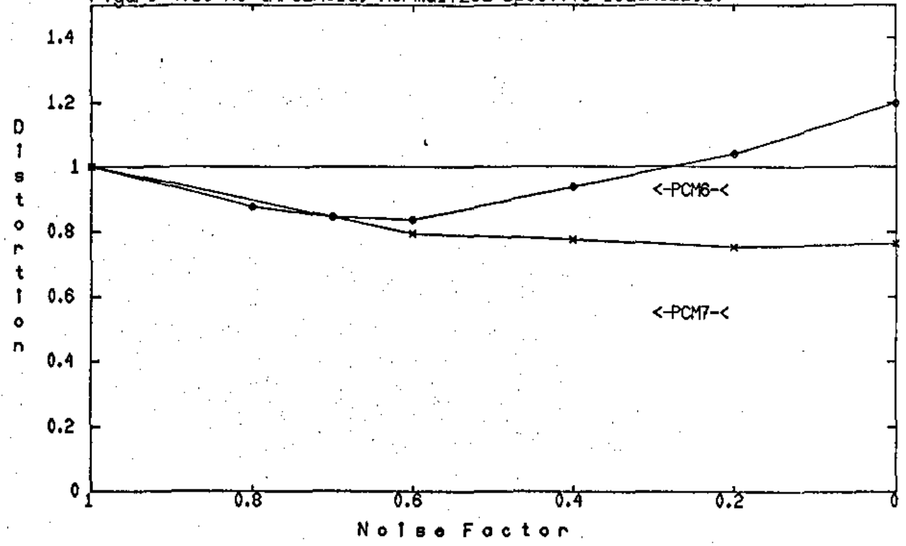


Figure 4.20 Difference function with threshold, no normalization.

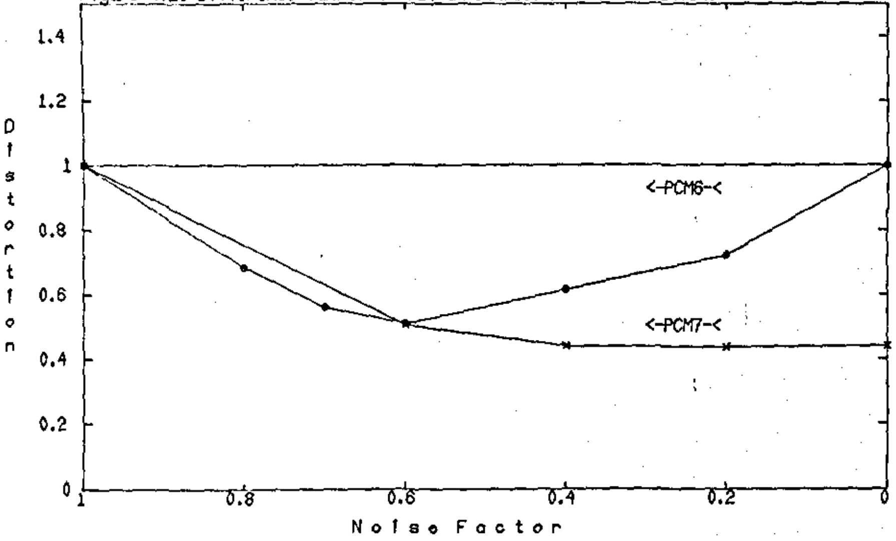


Figure 4.21 Difference function with threshold and SL normalization.

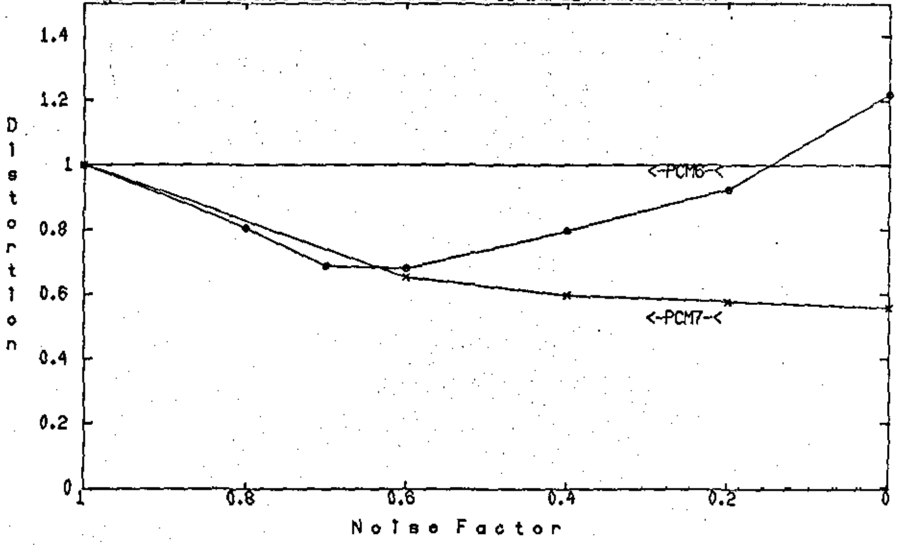


Figure 4.22 160 msec on loudness, no threshold, no normalization.

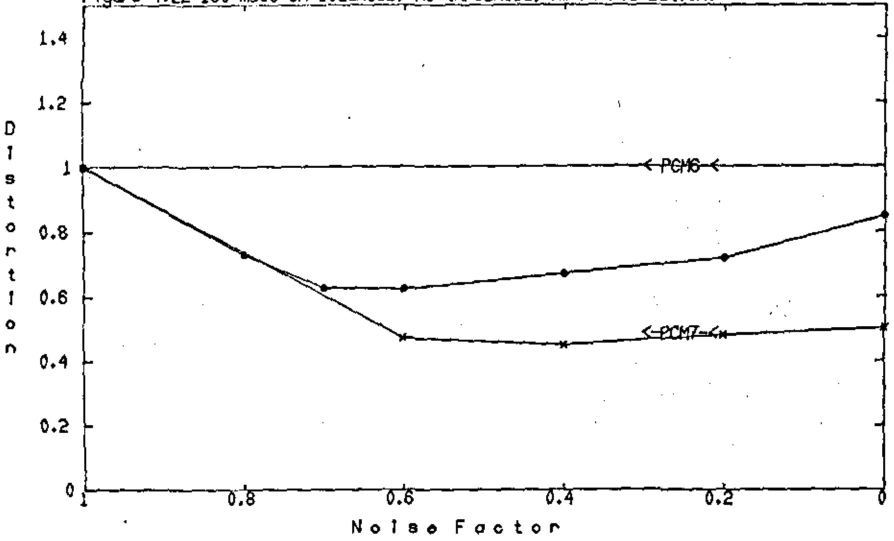


Figure 4.23 160 msec on loudness, no threshold, normalized.

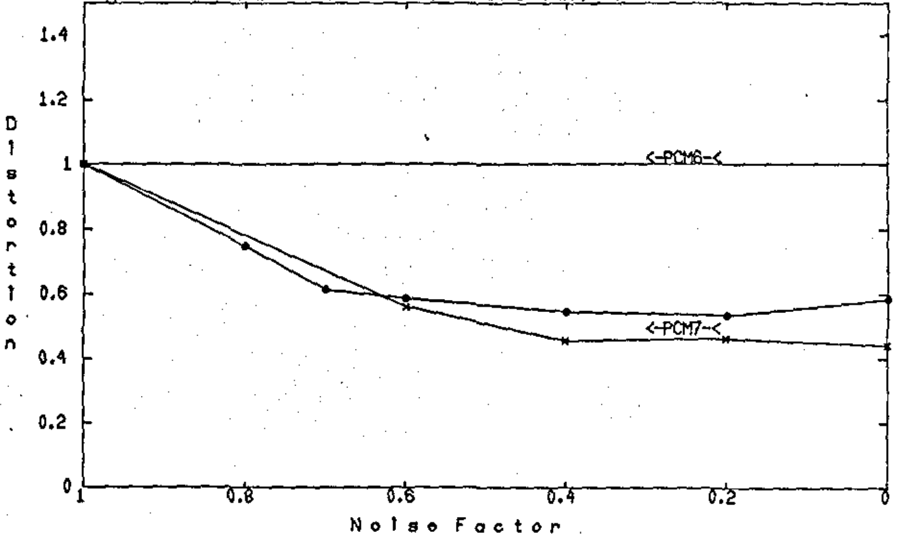


Figure 4.24 160 msec on loudness, with threshold and normalization.

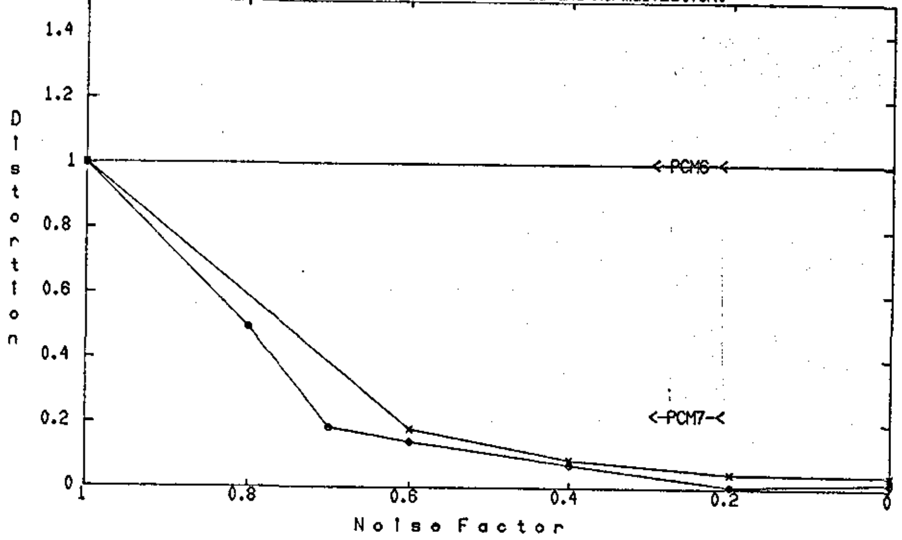


Figure 4.25 Ratio measure.

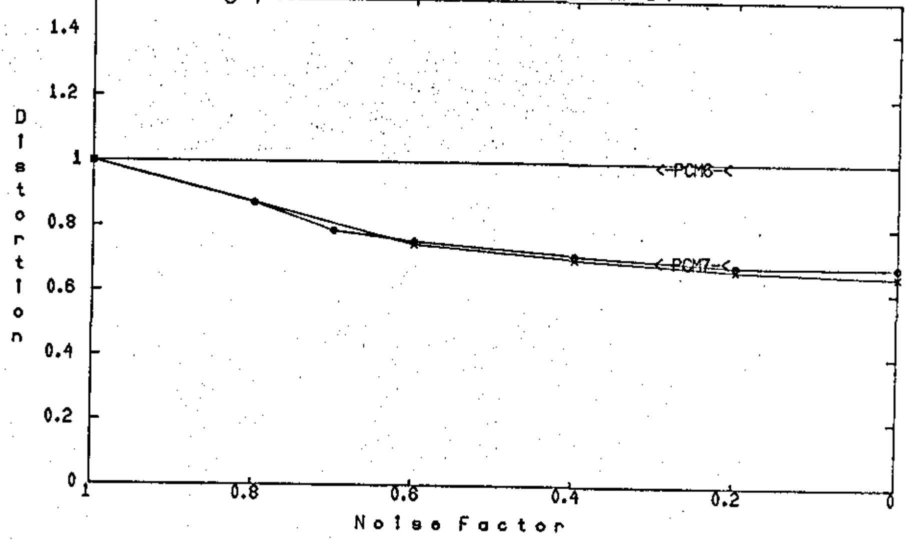


Figure 4.26 Log ratio measure.

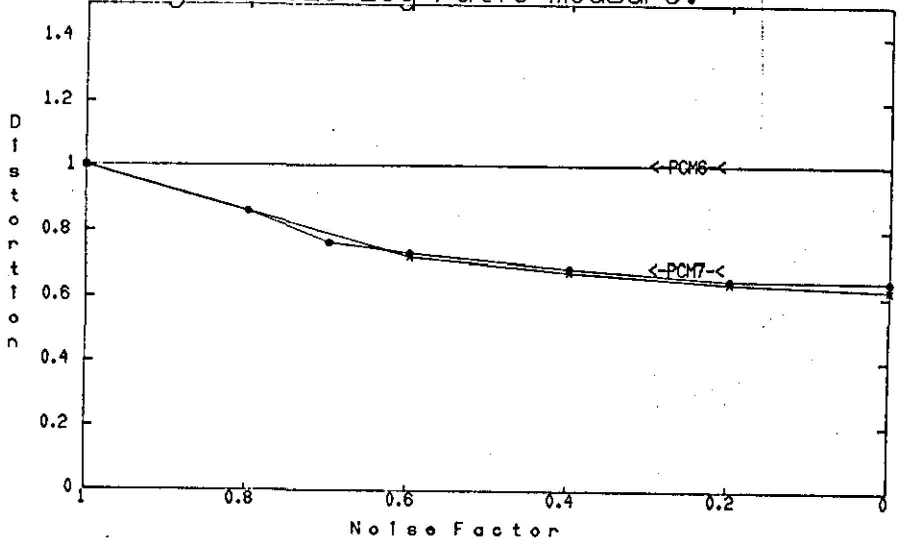


Figure 4.27 Variance measure.

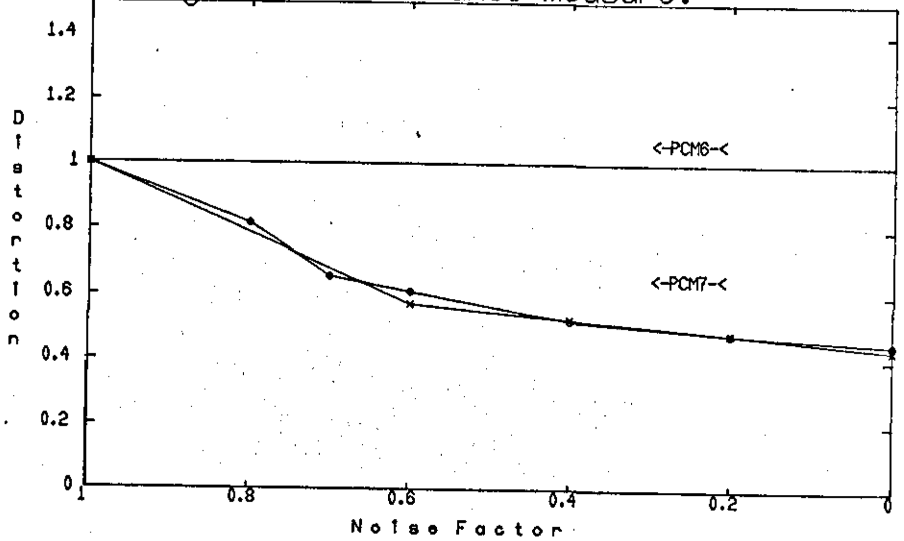


Figure 4.28 Normalized specific intensities, no threshold.

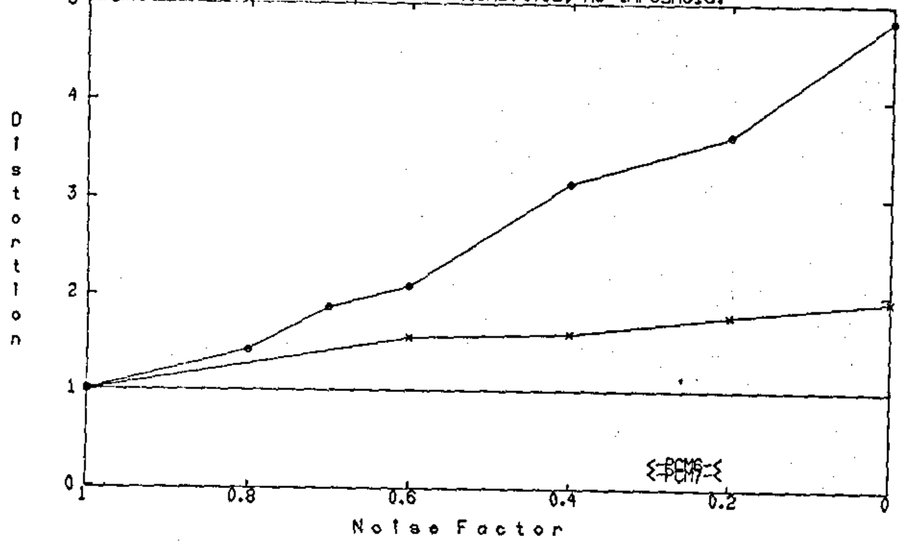


Figure 4.29 Equal bandwidth filters, no normalization.

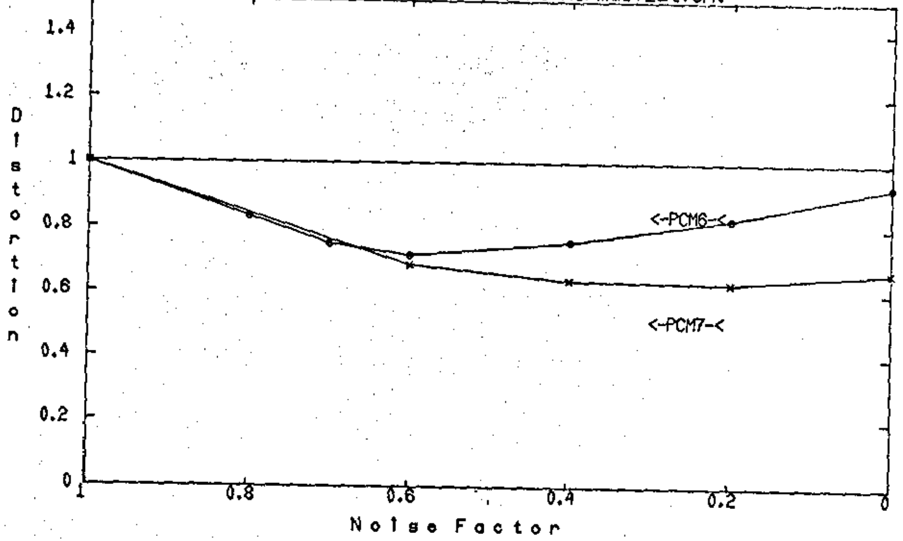
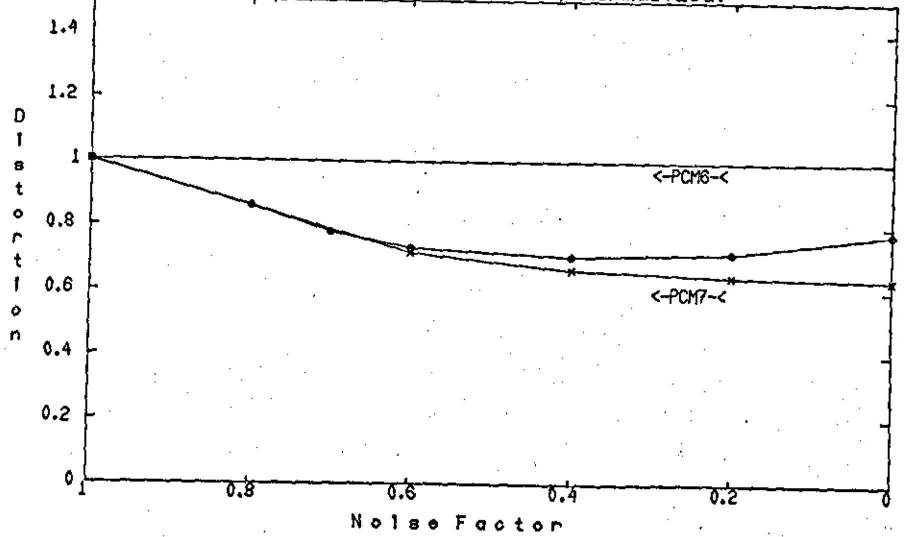
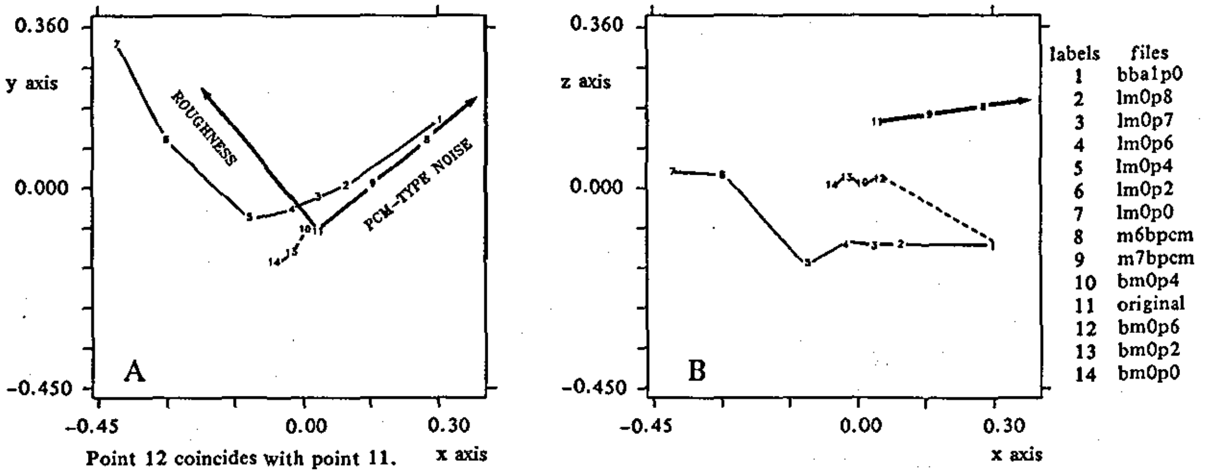


Figure 4.30 Equal bandwidth filters, normalized.





Point 12 coincides with point 11.

Figure 4.31 Difference algorithm, no threshold, no normalization.

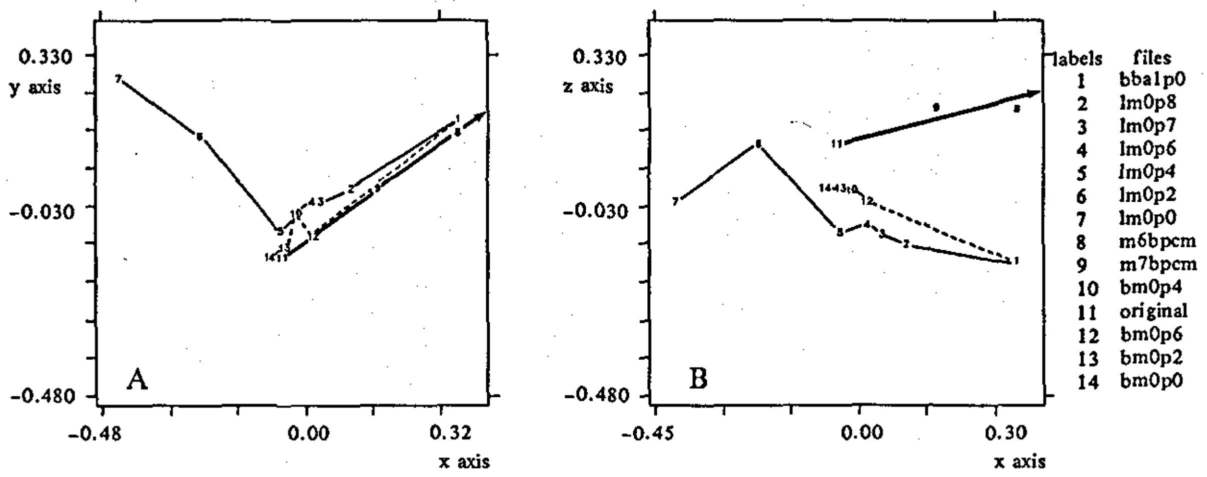


Figure 4.32 Difference algorithm, threshold included, no normalization.

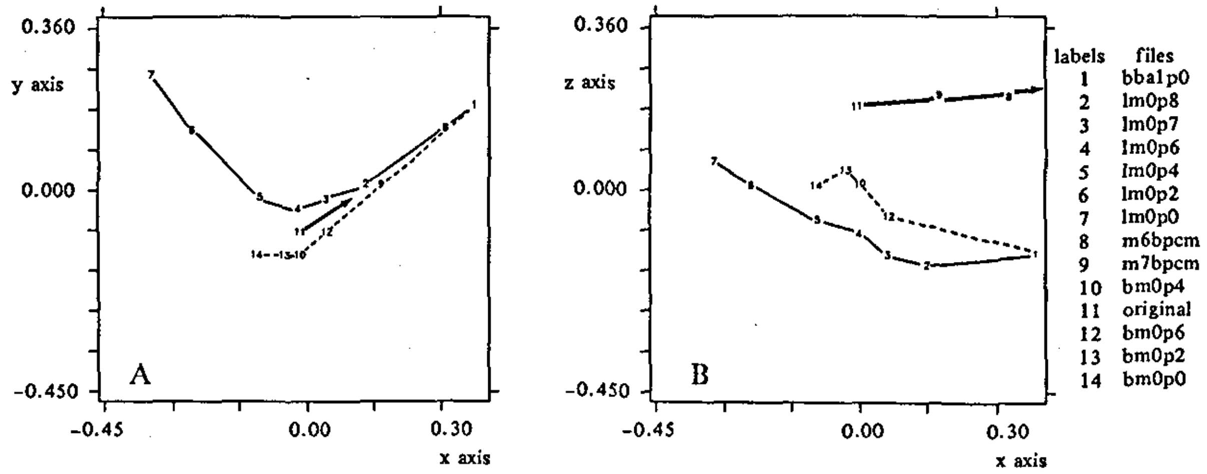


Figure 4.33 Equal bandwidths, no threshold, no normalization.

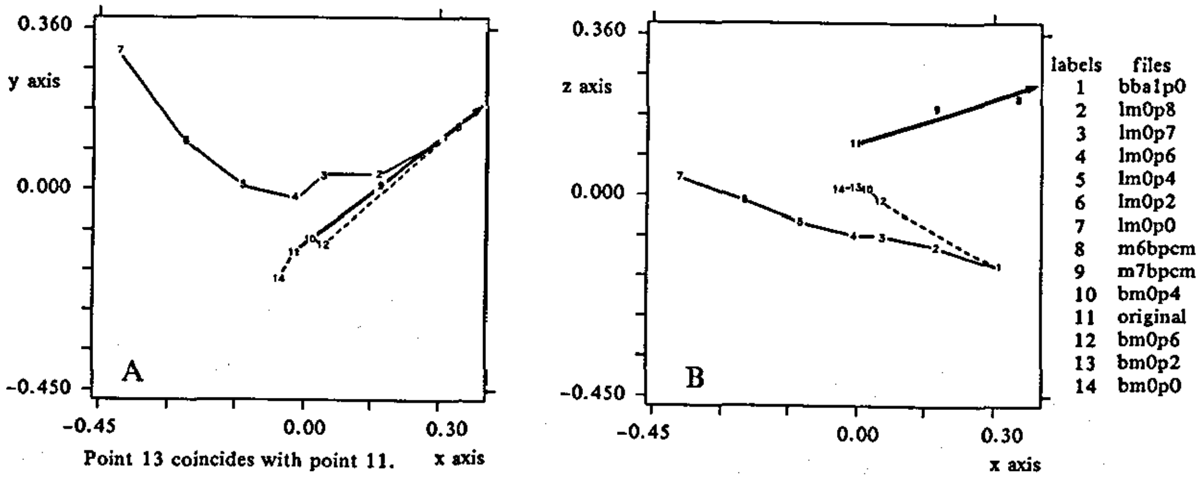


Figure 4.34 Difference algorithm, including threshold and normalization.

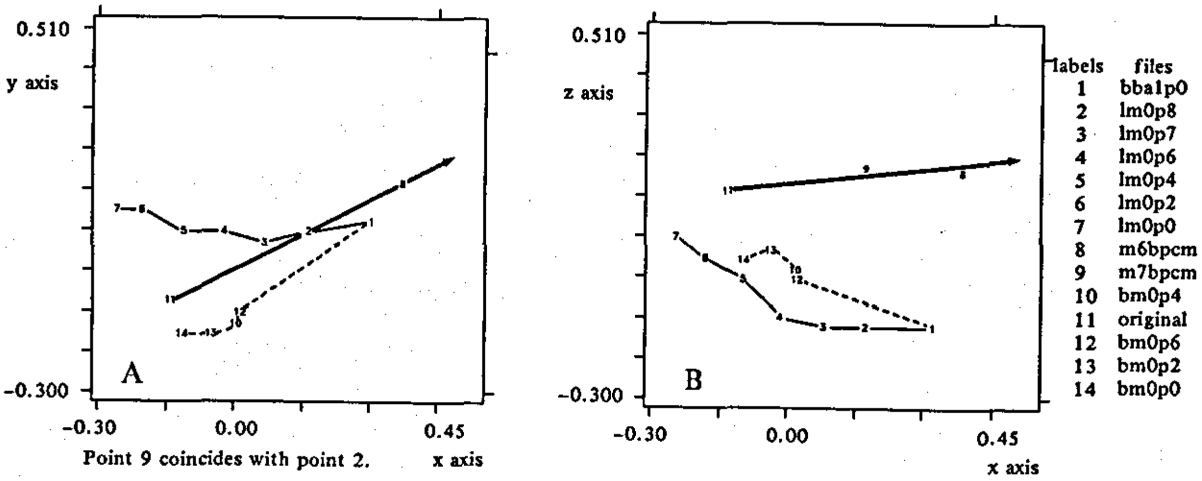


Figure 4.35 Log Ratio algoritm.

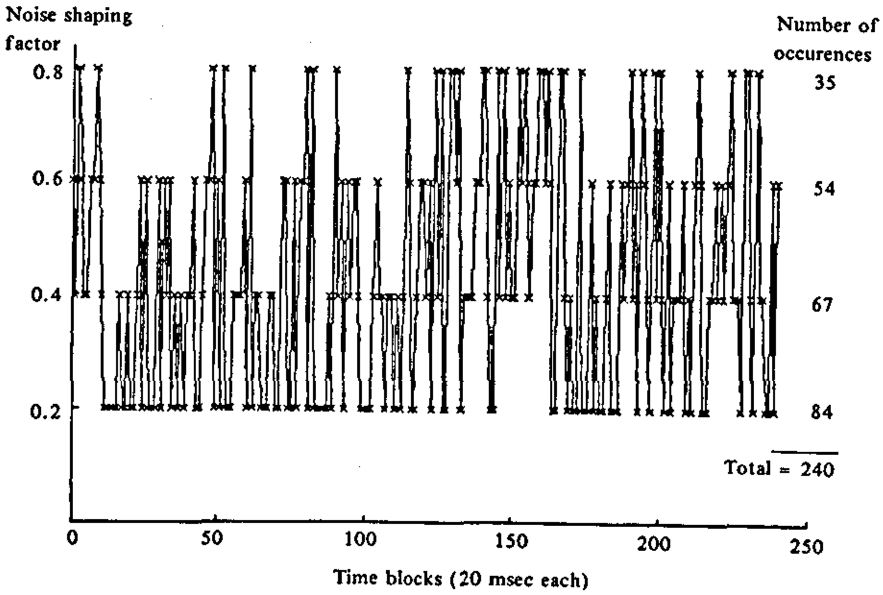
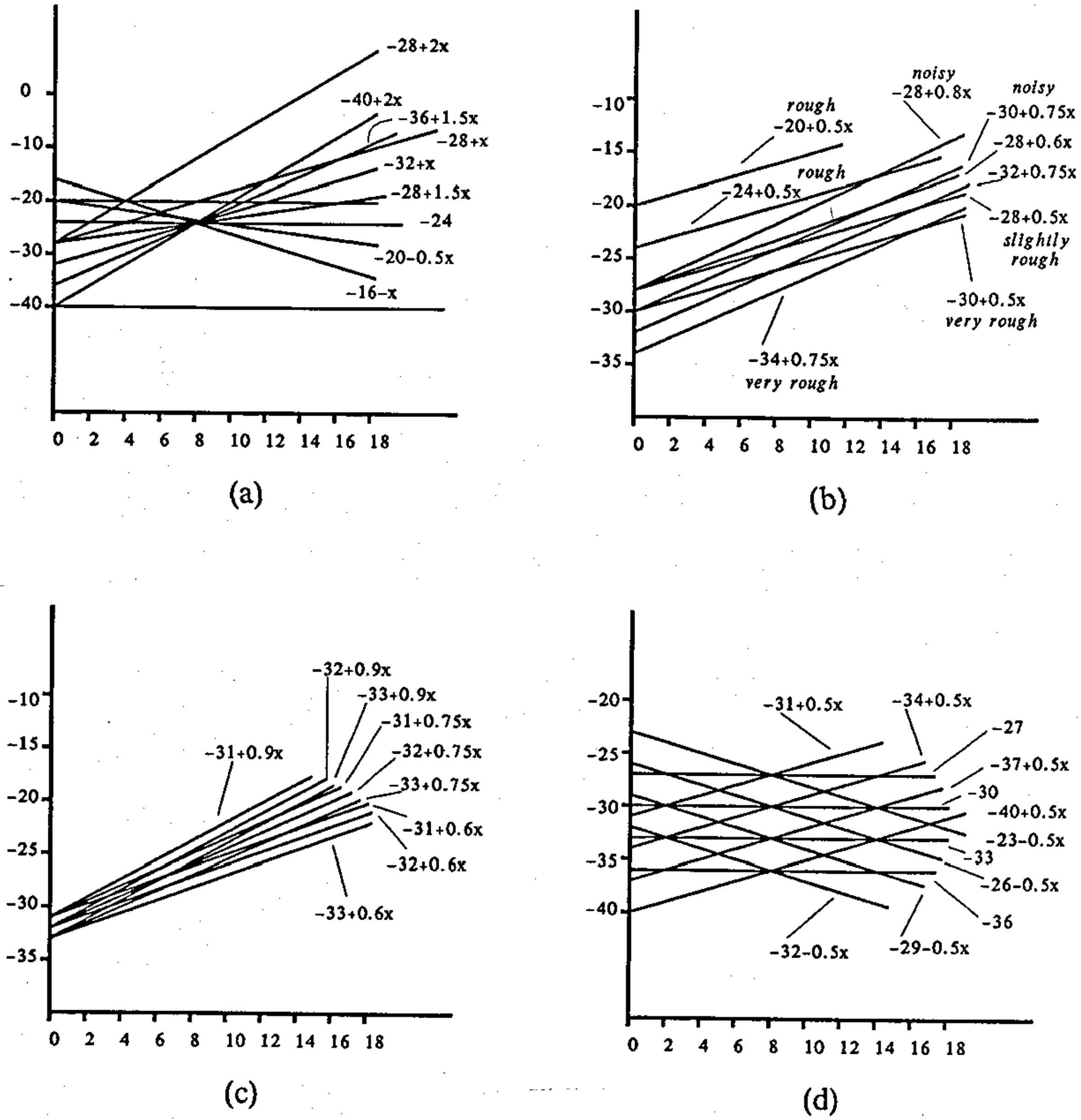


Figure 4.36 Noise factor chosen for each block. The number of occurences of each noise factor value is also shown on the right.



(a)

(b)

(c)

(d)

Figure 4.37 a-d Candidate threshold functions. Successively narrower ranges were explored around the most promising values. The performance of each function was assesed through subjective tests. The parameter selected was the noise shaping factor. An example of the perceptual attributes upon which the subjective assesment was based is shown in (b). For more details see text.

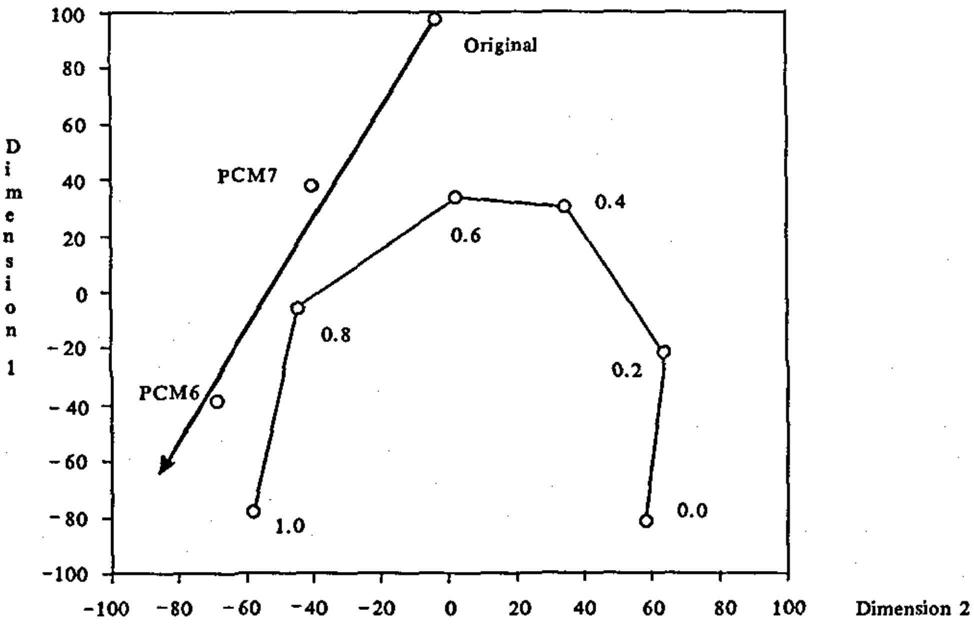


Figure 4.38 Final configuration obtained with the MINISSA algorithm. The subjective test results presented here refer to the same speech segment encoded with the ADPCM noise shaping algorithm. Different noise shaping factors are used to encode each segment. The noise shaping factors are shown as labels for each speech segment (point). The type and amount of distortion in each segment can be evaluated with reference to the original (uncoded) segment and the PCM axis defined by the two PCM coded segments, encoded with 6 and 7 bits per sample (PCM6 and PCM7 respectively).

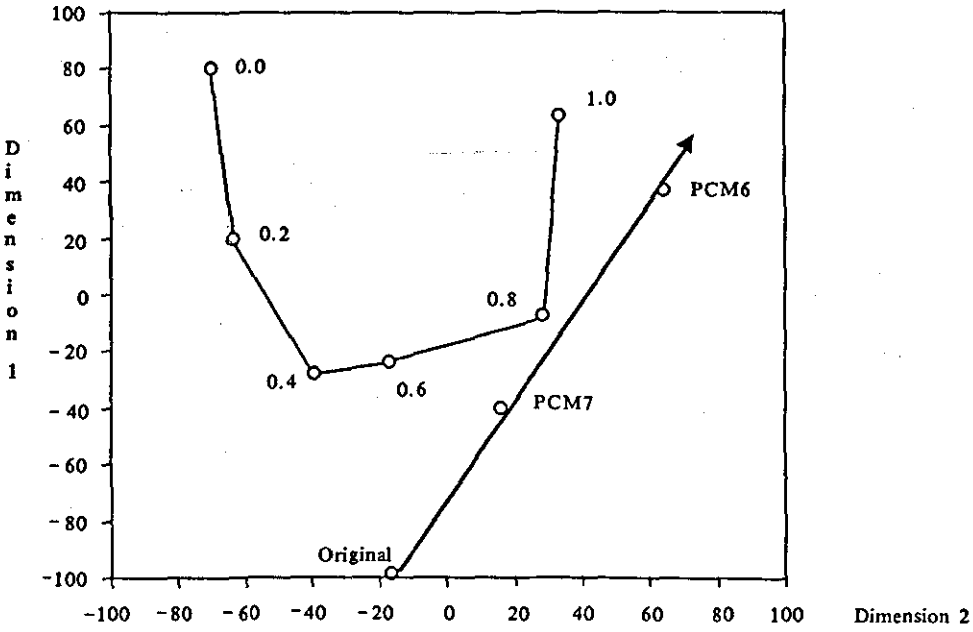


Figure 4.39 Final configuration obtained with the TRISOSCAL algorithm. The subjective test results presented here refer to the same speech segment encoded with the ADPCM noise shaping algorithm. Different noise shaping factors are used to encode each segment. The noise shaping factors are shown as labels for each speech segment (point). The type and amount of distortion in each segment can be evaluated with reference to the original (uncoded) segment and the PCM axis defined by the two PCM coded segments, encoded with 6 and 7 bits per sample (PCM6 and PCM7 respectively).

1. T. Berger, "Rate distortion theory" Prentice-Hall, Englewood Cliffs, N.J. 1971.

2. J.T. Rickard "New Fidelity Criteria for Discrete-Time source Encoding" IEEE Trans. on Information Theory vol IT-25 No 3, May 1979 pp275-282.

3. N.S. Jayant, P. Noll "Digital Coding of Waveforms" Prentice-Hall Inc. Englewood Cliffs, New Jersey 1984.

4. R.A. McDonald and P.M. Schultheiss "Information rates of Guassian signals under criteria constraing the error spectrum" Proc. IEEE, Vol 52, pp415-416, 1964.

5. M.R. Schroeder, B.S. Atal "Rate distortion theory and predictive coding "IEEE proc. ICASSP 1981 pp201-204.

6. R.E. Crochiere, L.A. Rabiner, N.S. Jayant, J.M. Tribolet "A study of Objective measures for speech Waveform Coders". Proc. Int. Zurich Seminar on Digital Communications ppH1.1-H1.7, March 1978.

7. J.M. Tribolet, P. Noll, B.J. McDermott, R.E. Crochiere "A study of complexity and Quality of Speech Waveform Coders" IEEE Proc ICASSP 1978 pp586-590.

8. B. McDermott, C. Scagliola, D. Goodman "Perceptual and Objective Evaluatiorn of Speech Processed by Adaptive Differential PCM" Bell System Technical Journal, May-June 1978 pp1597-1618.

9. P. Mermelstein "Evaluation of a segmental SNR measure as an indicator of the quality of ADPCM coded speech" J. Acoust. Soc. Am. 66(6) Dec. 1979 pp1664-1667

10. N.R. French and J.C. Steinberg "Factors Governing the Intelligibiity of Speech Sounds" J. Acoust. Soc. Am. vol. 19 number 1, Jan. 1947 pp90-119.

11. K.D. Kryter "Methods for the Calculation and Use of the Articulation Index" J. Acoust. Soc. Am. Vol. 34, No. 11, Nov. 1962 pp1689-1697.

12. K.D. kryter "Validation of the Articulation Index" J. Acoust. Soc. Am. Vol. 34, No. 11, Nov. 1962 pp1698-1702.

13. M.R. Schroeder, B.S. Atal, L. Hall "Optimizing digital speech coders by exploiting masking properties of the human ear" J.Acoust. Soc. am. 66(6) 1979, pp1647-1652.

14. N. Nocerino, F.K. Soong, L.R. Rabiner, D.H. Klatt "Comparative Study of Several Distortion Measures for Speech Recognition" IEEE Proc. ICASSP 1985 pp25-28.

15. A.H. Gray, J.D. Markel "Distance Measures for Speech Processing" IEEE trans on Acoust. Spech Sign. proc. Vol. ASSP-24 No 5, Oct. 1976, pp380-391.

16. R.M. Gray, A. Buzo, A.H. Gray, Y. Matsuyama "Distortion Measures for Speech Processing" IEEE trans. on Acoustics, Speech and Signal Proc. Vol ASSP-28 No 4. Au. 1980 pp367-376.

17. N. Kitawaki, K. Itoh, M. Honda, K. Kakehi, "Comparison of Objective Quality Measures for Voiceband Coders" ICASSP 1982, pp1000-1003.

18. J.L. Flanagan "Speech Analysis, Synthesis and Perception, Springer-Verlag-Berlin-Heidelberg-New York, 1972.

19. E. Zwicker "Procedure for calculating loudness of temporally variable sounds" J. Acoust. Soc. Am. Vol. 62, No 3, 1977, pp675-682.

20. E. Zwicker, E. Terhardt, E. Paulus "Automatic Speech recognition using psychoacoustic models" J. Acoust. Soc. Am. Vol 65(2), 1979 pp487-498.

21. F. Yeoh "Ph.D. Thesis" Loughborough University of Technology, 1983.

22. R.J. Harris "A primer of multivariate statistics" Academic Press, New York, San Francisco, London 1975.

23. Genstat V Mark 4.03, Lawes AGricultural Trust (Rothamsted Experimental Station)

24. J.C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis" Biometrica (1966) vol 53, 3 and 4, p325-338.

25. D.J. Goodman, R.D. Nash "Subjective Quality of the same Speech Transmission Conditions in Seven Different Countries" IEEE trans. on Comm. Vol. COM-30, No 4, April 1982, pp642-654.

26. J. Svean "Quality Measurements on Speech Coders for Mobile Radio" IEEE Proc ICASSP 1982, pp1700-1703.

27. W.D. Daumer "Subjective Evaluation of Several Efficient Speech Coders" IEEE trans on Comms. Vol. COM-30, No 4 April 1982, pp655-662.

28. W.D. Voiers "Diagnostic Acceptability Measure for Speech Communication Systems. IEEE proc. ICASSP 1977 pp204-207.

29. T.P. Barnwell III, S.R. Quackenbush "An Analysis of Objectively Computable Measures for Speech Quality Rating" ICASSP 1982 pp996-999.

30. V.R. Viswanathan, W.H. Russell, A.W.F. Huggins "Objective Speech Quality Evaluation of Medium band and Narrow band Real-time Speech Coders" ICASSP 1983 pp543-436.

31. S.R. Quackenbush and T.P. Barnwell III "Objective Estimation of Perceptually Specific Subjective Qualities" ICASSP 1985 pp419-422.

32. J.D. Carroll and M. Wish "Multidimensional Perceptual Models and measurement methods" in: Handbook of Peception Vol. II, Psychophysical Judgement and Measurement, E.C. Carterette and M.P. Friedman (Eds.) Academic Press, 1974 pp391-447.

33. M. Wish and J.D. Carroll "Applications of Individual Differences Scaling to studies of Human perception and Judgement" In Handbook of Perception Vol. II, E.C. Carterette and M.P. Friedman (Eds.) Academic Press, 1974, pp449-491.

34. T. Indow "Applications of Multidimensional Scaling in Perception" In Handbook of Perception Vol. II, E.C. Carterette and M.P. Friedman (Eds.) Academic Press, 1974 - pp493-531.

35. E.E. Roskam "The method of triads for Nonmetric Multidimensional Scaling" pp404-417.

36. B.J. McDermott "Multidimensional Analyses of Circuit Quality Judgements" J. Acoust. Soc. Am. Vol 45, No 3, 1969, pp774-781.

37. MDS(X) Programs: An Integrated Series of Multidimensional Scaling Programs with a Common Command Language. Originator: E.E. Roskam, University of Nijmegen, The Netherlands.

CHAPTER 5
VARIABLE RATE CODING

5.1 INTRODUCTION

Variable rate coding provides an attractive alternative to fixed rate coding for multi-user systems such as the telephone network. The output rate at the encoder can be made adaptive to both the single-user speech activity and the channel loading, as shown in figure 5.1. Speech activity can be defined in various ways. The simplest one is that of a non-silent situation: In normal conversation, the channel is occupied by a single user in only about 40% of the time, the other 60% of the time being silence. Speech activity for the purposes of silent-non silent classification can be measured over blocks of around 16ms for which the speech signal can be considered as stationary. For a single user, delays in the order of hundreds of milliseconds are necessary in order to utilize silence statistics [2,3] whilst in a multi-user situation, similar gains can be achieved over constant rate coding, even for delays of the order of the basic block of 16 ms [2,3]. Such gains (the so-called TASI advantage) are already utilized in systems such as TASI (Time assignment Speech Interpolation) and its digital counterpart DSI (Digital Speech Interpolation). More sophisticated approaches divide speech activity into more classes such as narrow-band voiced, wide-band voided, unvoiced and silence [4,5]. These systems are based on fairly sophisticated speech detectors which discriminate between the speech activity classes. Often, these detectors are also required to discriminate between speech and voiceband data [6], since data will normally require to be transmitted at the highest bit rate allowed by the network. A more formal approach to the problem is to employ rate distortion theory. The situation is virtually identical to that of subband coding where subband channels compete for bits, the total number of which is constrained by the overall bit rate. Here the "channels" can either be different users, or alternatively they can represent time blocks from the same user. In general a combination of both will probably be applicable. The resulting equations for optimum bit rates per "channel" are also similar to the subband coding situation and, likewise, the optimum solution strives to make the distortions (mse) from different channels equal in magnitude.

Rate distortion theory and its results should be applied with caution in variable rate coding. Although the advantage of variable rate is clear in the case of a silent-non silent situation, the same cannot be said for the within nonsilent segments situation. Firstly, there is a variation in level within speech from the same speaker and between speakers, which has no bearing upon variable rate and does not warrant different bit allocation for different levels: Clearly, the distortion itself should follow the variation in speaker level instead of being constant, as would be the case for "optimum" bit allocation dictated by rate distortion theory. Variations in average speaker level are measured over time intervals of the order of seconds. A second type of variation is syllabic variation: Different types of sounds are spoken with different intensities, with the vowels being at one end of the scale and the unvoiced fricatives at the other end of the scale. Individual variations also exist amongst different vowel sounds. Here again, perceptual studies indicate that the distortion should be proportional to the signal itself and not constant, as mse rate distortion theory (RDT) would have it. The principle of noise shaping is applicable here. One study [7] has shown that for the case of fairly high quality 16 kbs/sec coded speech, a bit allocation which is closer to constant bit alocation rather than a bit alocation given by RDT is preferred subjectively. Stationary blocks for syllabic variations such as the above can be of the order of tens of milliseconds up to about a quarter of a second. Within the above interval further variations exist, mainly due to the pitch structure of the speech signal. Such variations can be exploited quite successfully, [8,9] through RDT, to provide pitch related gains. Delayed decision coding algorithms especially pitch predictive algorithms inherently take advantage of this variation and no further gain can be expected from variable rate coding, within the above interval of 16-200 msec (depending on phoneme duration). This can be easily deduced by looking at the second residual which is nearly Gaussian in nature. (Delayed Decision Coding can in fact be considered as a variable coding scheme since, essentially block allocation of bits is effected, with implicit variable bit allocation within the block).

A source that outputs a variable rate bit stream presents a problem to channels that expect fixed rate data. In the case of multiplexing several users, all entering the network at the same node, there is no problem, since, the sum of the individual rates will be constant. In the case where the total source input to the channel is not constant, appropriate buffers and buffer control must be used to interface the variable source to the fixed rate channel. Several problems such as overflowing of the buffers emerge and various studies have been undertaken to produce useful systems for the network [3, 10, 11, 12, 13].

In the above situations variable rate coding (VRC) was used as a means of improving on speech quality by taking advantage of the nonstationarity of the source(s). VRC can also be dictated by the network itself depending on channel loading: During intervals where the load is low, a high rate can be afforded by individual users whereas under conditions of heavy load more severe bit rate constraints can be imposed upon users of the network. Two situations can be distinguished, one where the congested network node signals back to the individual encoders to reduce their output bit rates and the other where the note itself is able to "strip" bits off the transmitted codewords and reduce the bit rate for the subsequent paths in the network, without the need of signaling back to the encoder. The latter method necessitates coding structures (embedded coding) in which the bit rate reduction can be made possible after the encoding. This is usually achieved by providing one basic representation of the signal plus additional bits for refinements of the basic representation which can either be kept or stripped off according to available channel capacity [14, 15]. PCM systems inherenty possess this structure whereas embedded ADPCM structures have been deviced [14]. The available range of bit rates can extend from the vocoder range right up to the high quality waveform coder range, through the use of hybrid coders such as RELP [15] and subband coder/channel vocoder combinations [1]. A coder structure presented in chapter 6 seems particularly suitable for the above task.

5.2 VARIABLE RATE CODING SUBJECT TO A PERCEPTUAL CRITERION

5.2.1 "Real Time" Algorithms (suitable for speech transmission)

In order to assess further the capabilities of the distortion model developed in chapter 4 and to exploit the possibility of variable rate coding (VRC) subject to a perceptual criterion, a variable rate version of the multipulse coder ([16] appendix F) was developed. As mentioned briefly in chapter 3 the multipulse coder represents a speech segment by an all-pole filter and a number of pulses. The bit rate required to represent a given segment is therefore split into two parts: (a) The bits required for the quantization of the LPC coefficients and the pulse amplitudes, which can be done in a variety of ways as described in chapter 3. (b) The bits required to encode the locations of the pulses. For k pulses in a frame of L samples there are

$$C = \frac{L!}{(L-k)!k!} \qquad 5.1$$

possible sets (combinations). The minimum number of bits required to code the above positions is given by

$$C_b = \log_2 C \quad \text{bits} \qquad 5.2$$

One way to achieve the theoretical minimum given by 5.2 is enumerative coding [17].

For reasonable bit rates (around 9.6 kbs/sec) and full-band speech coding, only a small proportion of the L locations is occupied by pulses (i.e. k<<L). Typically k is about 10% of L. Around this percentage, the relationship between the number of pulses and the bit rate required for the coding of positions is very nearly linear. Further, for small variations of k, the bits required for the LPC coefficients and the amplitudes can remain constant and the bit rate can be varied by varying the number of pulses alone, through 5.2 and 5.1. For a bit rate around 9.6 kbs/sec, most of the mse distortion in the signal is introduced by the small k/L ratio. Only a small part of the m.s.e. distortion is introduced through the quantization of the LPC coefficients and amplitudes. The difference in SNRSEG between files with quantized and files with unquantized parameters (LPC coefficients and amplitudes) is often a fraction of 1 dB. Therefore, to remove any small effects of parameter quantization from this study, unquantized parameters were used in the simulations. For each time slice (frame) equation 4.43 can be modified into

$$d(t,k) = \sum_{z=1}^{n} [F(SL_o(z,t), SL_c(z,t,k)] \qquad 5.3$$

For each separate frame t, the distortion d can be represented as a function of the number of pulses k. The total distortion over a particular speech segment is then given by (as in 4.44)

$$D = \sum_{t=1}^{T} d(t,k) \qquad 5.4$$

in the case of a fixed rate coder, where the fixed rate is determined by k, the number of pulses.

For a variable rate coder equation 5.4 is simply replaced by

$$D = \sum_{t=1}^{T} d(t,k_t) \qquad 5.5$$

where, now, k_t is a function of the frame t. In order to minimize D in 5.5 it is necessary to know the relationship

$$d(k) = d(t,k_t) \qquad k=1,2....k_{max}$$

for each frame t. Unless otherwise specified, the function F in 5.3 is as given by equations 4.46-4.49. Even if the functions given by 5.6 are known for every t it is still difficult to minimize D in 5.5. An alternative approach, and the one adopted here, is to attempt to make d the same for each block, for a given average bit rate. The rationale behind this argument is that if d is a true perceptual measure of quality, then a constant perceptual distortion over time will be achieved, a desirable condition.

The criterion therefore is to minimize the deviation from a constant value of the distortion, d_{con}, i.e. minimize

$$d(t,\ k_t) - d_{con} \qquad 5.7$$

for all t, subject to the constraint of a constant average bit rate. Since a near linear relationship exists between bit rate and pulse rate, 5.7 is minimized subject to the constraint of a constant average pulse rate. The TASI advantage is of no interest here. Therefore, silent segments are coded at the average bit rate (fixed value) and do not compete for pulses with non-silent segments. Two sentences from a male speaker were used for the simulations, the same ones as the ones used in chapter 4: "There was an old man called Michael Finnegan, he grew whiskers on his chinnagen". The frame size L was fixed at 120 samples and two different target average bit rates were used, one at k = 12 pulses/frame and the other at k = 7 pulses/frame. The minimum number of pulses in a frame was set to one, although a frame with zero pulse allocation can also be conceived, in the situation where the ringing (filter memory) from the previous frame is sufficient for the reproduction of the current frame. The maximum number of pulses was set to 24. Although no upper limit is necessary, the above value sets the higher target value for the average bit rate in the middle of the range. An unlimited range of pulses is unpractical since, for each value of K, the value for d(k) from 5.6 need be calculated. Note that the iterative nature of the multipulse solution provides all the synthesized frames for $k=1,2....k_{max}-1$, for a final value of $k=k_{max}$ without any additional computation.

As a first approximation, the assumption was made that equation 5.6 was of the form

$$d(k) = a/k^r \qquad k=1,2,.....k_{max} \qquad 5.8a$$

or

$$d(t,k_t) = a_t/(k_t)^{r_t} \qquad k=1,2....k_{max} \qquad 5.8b$$

therefore, in logarithmic terms a linear relationship obtains:

$$\log d(t,k_t) = \log a_t - r_t \log(k_t) \qquad k=1,2,....k_{max} \qquad 5.8c$$

between the logarithm of distortion and the logarithm of the pulse rate, for a particular frame, t.

The validity of the assumption can be tested by plotting d(k) for each pulse rate k for different frames. Several such plots for successive frames are shown in figure 5.2. It can be seen that to a first approximation, the function is indeed a straight line as given by 5.8c. The superimposed straight lines in figure 5.2. are obtained by evaluating $\log a_t$ and r_t through a least squares procedure for straight line fitting. Although the general trend is a straight line, deviations can be seen for various frames. In particular the function d(k) is nonmonotonic at times. This may seem peculiar at first since it implies that the distortion actually increases with the addition of more pulses. The paradox is explained by the fact that the criterion optimized by the selection of pulses in the multipulse LPC is the weighted SNR, whereas the criterion actually measured is the one given by 4.46-4.49. Thus, although the distortion in terms of weighted SNR decreases with the addition of more pulses, d(k) can actually increase in certain situations. This generates complications manifesting themselves as local maxima and minima on the plots of figure 5.2. Effective ways to deal with these irregularities of d(k) will be described later. For the time being the relationship of d upon k is assumed to be given by 5.8c, where the constants $\log a_t$ and r_t are evaluated for each frame from data such as the ones shown in figure 5.2, through least squares procedures. This considerably simplifies the analysis since a monotonic relationship between d and k is known for each frame t. Furthermore, a straight line can be defined by only two points. Therefore, the possibility exists, that d(k) need only be evaluated at two points k_1 and k_2, to yield $d(k_1)$ and $d(k_2)$, instead of evaluating d at each $k=1,2,\ldots.k_{max}$, to obtain the function d(k). This would significantly reduce complexity. To summarize, for each block t, a relationship of the form given by 5.8 is assumed, where the constants are evaluated once for each frame as described above. Now assume that the reverse problem is addressed: Given a constant distortion d_{con}, find the required pulse rate k_t such that the following equation is satisfied:

$$d_{con} = a_t/(k_t)^{r_t} \qquad 5.9a$$

the solution is of course given by

$$k_t = (a_t/d_{con})^{1/r_t} \qquad 5.9b$$

To derive equation 5.9b, the relationship in 5.8a is assumed to be continuous, i.e. k_t is assumed to take noninteger values as well. Since the constants in 5.9b are known for each frame, the appropriate pulse rate can be derived. The average pulse rate k_{av} is given by

$$k_{av} = \sum_{t=1}^{T} k_t = \sum_{t=1}^{T} (a_t/d_{con})^{1/r_t} \qquad 5.10$$

Note that k_t as given by 5.9b can be outside the set limits of 1 and 24. In this situation, k_t is set to the value at the limit, i.e.

$$k_t = (a_t/d_{con})^{1/r_t} \quad , \quad 24 \geqslant k_t \geqslant 1 \qquad 5.11a$$

$$k_t = 24 \qquad \left(\frac{a_t}{d_{con}}\right)^{1/r_t} > 24$$

5.11b

$$k_t = 1 \qquad \left(\frac{a_t}{d_{con}}\right)^{1/r_t} < 1$$

5.11c

With the redefined values for k_t given by 5.11 a plot of k_{av} versus d_{con} can be plotted from the values of a_t and r_t alone. This plot is shown in figure 5.3.

Having obtained the relationship shown in figure 5.3 one can now reverse the step, to evaluate the required d_{con} for a preselected average pulse rate. For example, if an average pulse rate of 12 pulses/frame is required, this can be achieved by setting d_{con} to about 17. With the value of d_{con} known, the necessary pulse rate for each frame can be found from equations 5.11, the value rounded to the nearest integer. The average bit rate would then be close to the chosen value, in this case, 12 pulses/frame.

The graph of figure 5.3 is not expected to vary considerably amongst speakers. Therefore an equivalent graph can be derived for a large set of training data for different speakers. Subject to the assumption of a straight line relationship between the logarithms of distortion and pulse rate as given by equation 5.8c and to a certain degree verified in figure 5.2, a variable rate coding algorithm has been designed which offers the choice of a wide range of preselected average bit rates and strives to produce a perceptual distortion which is constant throughout the utterance, or even throughout the utterances of several speakers.

A schematic of the algorithm is shown in figure 5.4. Assume that the appropriate pulse rate has been selected for block t-1. The filter memory from block t-1 is then used in the calculations for block t. As explained above, for a maximum possible number of pulses equal to k_{max}, a number of k_{max} different multipulse frames are synthesized, giving a one pulse representation of the frame, a two pulse representation and so on, until a k_{max} pulse representation is obtained. The distortion d(t,k) for the current frame is calculated from equations 5.3 and 4.47-4.49 for each value of k = 1,2,....k_{max}. From these values, the constants $\log a_t$ and r_t are evaluated as described above. These two values define the appropriate straight line for block t. For a preselected average pulse rate k_{av} a corresponding constant distortion value is defined from equation 5.11. The intersection of the line d(t,k) = d_{con} (horizontal) with the line given by 5.8c gives the point whose abscissa is the selected value k_t. (After appropriate logarithmic transformations as dictated by the formulas). The selected value is then rounded to the nearest integer. The multipulse frame synthesized with the selected number of pulses k_t forms the current block and the algorithm proceeds to calculate the pulse rate and the multipulse frame for the next block of speech.

The algorithm was used to obtain a coded file with an average bit rate of 12 pulses/frame. The pdf of the distortion obtained is given in figure 5.5a. It can be seen that the algorithm has been successful in producing a near constant distortion, of value around 17.5. The average bit rate was indeed very close to 12 pulses/frame. The speech quality obtained though was considerably worse than the quality from a fixed rate coder operating at 12 pulses/frame. The pdf plot of the pulse rate shown in figure 5.5b gives a clue as to why this is so: The algorithm has assigned 24 pulses to a great number of frames which necessarily forces the pulse rate for the rest of the frames to quite low values which result in poor quality speech. This behaviour of the algorithm is not due to an inappropriate distortion measure but due to the straight line approximation adopted. This can be seen from figures 5.6-5.8 which show three representative plots of distortion versus pulse rate (both on logarithmic scales) for three different frames. The horizontal line gives the value of constant distortion at $\log_{10}d_{con} = \log_{10} 17.5 \approx 1.24$. It is clear from figures 5.6-5.8 that the straight line of $d_{con} = 17.5$ does not intersect the distortion line of equation 5.8c and therefore the condition of equation 5.11b is reached. The selected pulse rate is therefore 24 which is shown by the rhombus markers in figures 5.6-5.8. It is also clear from figures 5.6-5.8 that a great reduction in pulse rate can be achieved with a very small increase in distortion (or none at all) if the pulse rates shown by the octahedron-in-square (OIS) markers are chosen instead. This is because the rate of decrease of distortion in going from the OIS marker to the rhombus is very small. This is not typical of all the frames though and therefore some automatic procedure needs to be devised so that a more appropriate pulse rate can be chosen for each frame. For this purpose, the algorithm shown schematically in figure 5.9 was designed. Central to this algorithm is a threshold value, d_s, which is small by definition and its exact value is determined from subjective tests. The algorithm starts from an initial pulse rate which, in this case, is the value chosen from the straight line approximation. Let this pulse rate be given by k. The distortion d(k) is known. The algorithm compares d(k) with d(1). If $d(1) - d(k) < d_s$ then the chosen pulse rate is 1, otherwise d(k) is compared with d(2). If the difference is smaller than d_s then the chosen pulse rate is 2 and so on until an appropriate rate is found, or the possible values of k are exhausted. Note that the algorithm avoids local maxima because d(k) is compared with distortions of an ascending sequence of pulse rates. This is the single update largest jump (SULJ) algorithm. Its purpose is to reduce the cost (the pulse rate) with a minimal increase in distortion. The positions of the OIS markers in figures 5.6-5.8 have actually been determined through the above algorithm. It can be seen that the algorithm makes reasonable choices for the new pulse rates. For a value of d_s equal to 2.5, the pdf of the distortion of the speech file coded with the combination of the straight line algorithm followed by the SULJ algorithm is shown in figure 5.10a. By comparison to figure 5.5a, there is little change in the two pdf as would be expected since the maximum (single frame) change is constrained to be less than $d_s = 2.5$. The resulting speech quality is almost indistinguishable form the one obtained from the straight-line (SL) algorithm (in fact this was the criterion for choosing $d_s = 2.5$) but the resulting average pulse rate is about 7.5 pulses/frame! The pulse pdf for this configuration is shown in figure 5.10b. A reduction in pulse rate of almost a half is obtained with very little increase in distortion. In order to bring the average pulse rate back to its desired value of 12 pulses/frame a new algorithm was designed. Intuitively such an algorithm must reduce the distortion as much as possible with the smallest increase in pulse rate. This can be done by taking advantage of the individual shapes of each distortion versus pulse rate function for each block: Figures 5.11-5.14 show the choice of pulse rate made by the SL logarithm with a rhombus-in-square (RIS) marker. As before, this point is the intersection of the straight- line distortion curve approximation of equation 5.8c with the horizontal line $d = d_{csn}$. The chosen value is in accordance with equation 5.11a. The extreme situation of equation 5.11c is shown in figures 5.15-5.18. The octahedron markers in figures 5.11-5.18 show alternative pulse rates which are not very different from the (RIS) rates (remember that the logarithms of the pulse rates are shown on the abscissa) but for which the distortion is significantly reduced. These new locations were obtained through a new algorithm shown schematically in figure 5.19. This is the multiple update smallest jump (MUSJ) algorithm. It operates as follows: The distortion of the current pulse rate d(k) is compared with the distortion of the next pulse rate d(k+1). If the difference d(k) - d(k+1) is larger than a threshold value d_L then the new pulse rate k+1 is retained and the test repeated by comparing d(k+1) - d(k+2) with d_L. If the test is negative the old pulse rate is retained but the test now involves a comparison of d(k+1) - d(k+3) with d_L and so on until the maximum value k_{max} is reached. The formulation of the algorithm is designed to deal with local distortion maxima and aims to find a global distortion minimum subject to a small increase in pulse rate. The value for the threshold d_L can be determined from the final (required) bit rate. For the present example, in order to bring the bit rate from a value of 7.5 to the required average value of 12 pulses/frame a value for d_L = 2.3 was used. The resulting pdf of distortion is shown in figure 5.20a. It can be seen that the distortion for the coded file is less than the distortion shown in figure 5.5a for the same average pulse rate (12 pulses/frame). Also, the distortion variance in figure 5.20a is larger than the distortion variance in figure 5.5a. This was to be expected since the SL algorithm whose results are shown in figure 5.5a aims to minimize this variance. Therefore the minimization of distortion and a flat distortion over time are not equivalent criteria. This is a direct consequence of the irregular shape of the distortion versus pulse rate functions. In turn, this is due to the fact that the multipulse algorithm reduces the weighted m.s.e. by each additional pulse and not the perceptual distortion measure of equation 5.3.

The resulting pdf of the pulse rate obtained by successive application of the SL, SULJ and MUSJ algorithms is shown in figure 5.20b. Note that this pdf is free from the accumulation of pulse rates at the ends of the pulse rate range. The successive operations of the three combined algorithms are also shown in four typical plots of distortion versus pulse rate for four different frames, in figures 5.21-5.24. The rhombus is the pulse rate chosen by the SL algorithm, the square the subsequent correction from the SULJ algorithm and, finally, the octahedron the last adjustment of the MUSJ algorithm. The final, combined, algorithm gave a quality which was indistinguishable from the file coded at the fixed rate of 12 pulses/frame. Note that since the combined algorithm has three "free" parameters and only one constraint (the average pulse rate), various combinations of d_{con}, d_{s} and d_{L} can produce an average pulse rate of 12 pulses/frame. Several combinations were attempted but the subjective quality could not be improved over the quality of the fixed rate file. It is quite remarkable that a pulse distribution such as in figure 5.20b yields a file which sounds exactly the same in quality as the one with a fixed rate. A similar situation was encoüntered for a target rate of 7 pulses/frame. The results of using the SL algorithm with a target average bit rate of 7 pulses/frame are shown in figures 5.25 a,b with $d_{con} \simeq 30$ as can be deduced from figure 5.3. With the subsequent use of the SULJ algorithm the figures 5.26a,b were obtained. The average pulse rate dropped to 5.2 pulses/frame with $d_{s} = 0.0$. A reduction in pulse rate occurs even with a zero threshold because of the nonmonotonic nature of the individual distortion versus pulse rate functions. Finally, the composite SL, SULJ, MUSJ algorithm was applied to obtain an average pulse rate/frame of 7 pulses/frame. The resulting pdfs are shown in figures 5.27a,b. The chosen values for d_{con}, d_S and d_L were not very appropriate in this case as can be seen by comparing the pdf of distortion of figure 5.27a with the pdf of distortion for the fixed rate coded speech (7 pulses/frame) of figure 5.28. Although the variable rate scheme reduces the large distortion values this is at the expense of a shift of the whole pdf towards higher values. A more successful selection was obtained by constraining k_{max} to a value of 12 (so that the average of 7 is in the middle of the allowable range), fixing d_{con} at a very high value (i.e. $d_{con} = \infty$) so that the SL algorithm, sets the pulse rates for all the frames to $k = 1$, a value for $d_S = 0$ and, finally, $d_L = 4$. The resulting average rate is again 7 pulses/frame. The corresponding pdfs are shown in figures 5.29 a,b. The pdf of the distortion is very similar to the pdf for the fixed rate coder although for the variable rate scheme the pdf is shifted slightly towards lower values and the bulge around $d \simeq 36$ in the fixed rate case is not present in the variable case. The combination of thresholds given above gave the best subjective performance, but again, this was very similar to the fixed rate performance.

The variable rate algorithms described above provide a suitable framework for determining appropriate values for parameters used in the distortion measures. In particular, the algorithm was used to determine appropriate values for the threshold parameters a,b in 4.48 and 4.66. Several combinations were used, shown in figure 5.30. Variable rate coded files were then produced with the values shown in fig. 5.30 applied into the distortion measure of 5.3. Through informal listening tests the best values, leading to a minimal distortion coded file were again as given by equation 4.49.

5.2.2 "Non-real time" algorithm (suitable for speech storage

In all of the above algorithms a decision for an appropriate pulse rate for a specific block depended on the distortion curve, $d_t(k)$, of that block alone. The information about the rest of the speech utterance, necessary to determine appropriate thresholds for the operation of the algorithms was provided in an initial training phase. Therefore the schemes presented above could be used in a real-time application.

In situations where the real-time restriction can be avoided, as in voice storage applications, more efficient algorithms can be employed. In these schemes the speech segment of interest is first stored in memory at a high bit rate, high quality mode which can then be reduced to a low bit rate mode, as a background task. The multipulse algorithm is particulary suited to such applications since a high pulse rate version of speech can be stored initially and then pulses "removed" on a frame to frame basis subject to a meaningful criterion, in order to obtain a lower bit rate, high quality speech signal. In this case the whole speech utterance and thus the curves $d(t,k_t)$ $t = 1,2....T$ are all known a priori. Several algorithms can be developed to produce a minimum distortion variable rate file using this a priori information. The two algorithms developed here represent some form of "steepest descent (or ascent)" algorithms. They differ mainly in the initial conditions assumed.

The first algorithm sets the pulse rate for each frame to the target average pulse rate. This is step 1. In step 2 the difference $d(t,k_t-1) - d(t,k_t)$ is formed for each frame and the frame t_{min} for which

$$d_{min} = d(t,k_t-1) - d(t,k_t) \qquad 5.12$$

is the smallest is found.

In step 3 the difference $d(t,k_t) - d(t,k_t+1)$ is formed for each frame and the frame t_{max} for which

$$d_{max} = d(t,k_t) - d(t,k_t+1) \qquad 5.13$$

is the largest and, in addition, $t_{max} \neq t_{min}$ is found. Step 4 compares the two differences: If d_{max} is greater than d_{min} then the pulse rate for t_{min} is reduced by one and the pulse rate for t_{max} is increased by one i.e.

$$\text{iff } d_{max} > d_{min} \text{ then } \left\{ \begin{array}{l} k_{tmin} \rightarrow k_{tmin}-1 \\ k_{tmax} \rightarrow k_{tmax}+1 \end{array} \right. \qquad 5.14$$

The above is an "exchange" algorithm in that the average pulse rate is held constant at any stage in the algorithm: Any increase in the pulse number in one frame is immediately compensated by a decrease in the pulse number in another frame. The algorithm, after step 4, returns to step 2 for a prespecified number of iterations or until the condition in 5.14 can no longer be fulfilled for any pair of frames. This algorithm is susceptible to local maxima and minima. One way to overcome this problem is to produce an alternative algorithm, such that 5.12 is replaced by

$$d_{min} = d(t,k_t-a) - d(t,k_t) \qquad 5.15$$

equation 5.13 is replaced by

$$d_{max} = d(t,k_t) - d(t,k_t+a) \qquad 5.16$$

and finally, test 5.14 replaced by

$$\text{iff } d_{max} > d_{min} \text{ then } \begin{array}{l} k_{tmin} \rightarrow k_{tmin}-a \\ k_{tmax} \rightarrow k_{tmax}+a \end{array} \qquad 5.17$$

and the algorithm run as above. The original algorithm is first applied until equation 5.14 cannot be fulfilled and then the modified algorithm is initiated (from step 2) for various values of a. When 5.17 can no longer be fulfilled, the first algorithm is then reapplied and so on. It was found that with a = 2, around five exchanges between the original and modified algorithms were enough to reach the situation where 5.14 or 5.17 were not satisfied, even from the first iteration. The number of iterations reduces very quickly as the number of exchanges between the original and modified algorithms is incresed.

It was found that this scheme produced results very similar in subjective quality to the other ("real-time") schemes and to the fixed rate algorithms. However, the pulse pdf had a much smaller spread than the other variable rate schemes described above and a triangular of gaussian shape as opposed to the more uniform shape of the "real-time" schemes (as in fig. 5.29b and 5.10b). The results were halfway between the fixed rate pdf and the pdf distributions of figures 5.29b and 5.20b. It seems that the exchange nature of the algorithm did not allow sufficient variation in the individual pulse rates, resulting in a narrow pdf.

Another algorithm, which still uses a priori information as above, but avoids the exchange mode restrictions that force the bit rate to be fixed to the average (target) value at each iteration, was developed. This algorithm starts with the assumption that any frames coded with the maximum value of pulses, k_{max}, have "negligible" (or acceptable) distortion. This is verified from subjective results where, if all the frames are assigned a pulse number $k = k_{max}$, very little distortion is perceived, which can be neglected when compared with the distortion produced at the target pulse rate.

At step 1 the algorithm sets the pulse number for each frame to the maximum value k_{max}. At step 2 the difference $d(t,k_t-1)-d(t,k_t)$ is formed and the frame t_{min} for which

$$d_{min}=d(t,k_t-1)-d(t,k_t) \qquad 5.18$$

is the smallest is found. For this frame, the pulse number is reduced by one and the algorithm returns to step 2 until the target rate is reached. Modified algorithms can also be used where the decrease in pulse rate is done at larger steps, to avoid local maxima/minima. This algorithm also gave subjective results similar to the fixed rate but, in addition the pulse (and distortion) pdf were very similar to the pdf of figures 5.20 and 5.29. The results for a target rate of 12 pulses/frame are shown in figure 5.31a,b.

5.3 Variable rate coding with a w.m.s.e. criterion

The multipulse algorithm [16] (appendix F) minimizes the weighted mean square error (wmse) $e_w^T e_w$ where

$$e_w = Y_w - X_w \qquad 5.19$$

with Y_w being the weighted input speech signal vector (minus the memory from the previous frame) and X_w the weighted, synthesized signal due to the pulses in the current frame alone. The weighting is effected through linear filtering with a transfer function

$$W = [1-P(Z)]/[1-P(Z/a)] \qquad 5.20$$

where P(Z) is the all pole predictor matched to the signal segment Y. This form of weighting is similar to the noise shaping operation in noise feedback coding (chapter 3). The wmse can be considered a function of the number of pulses in the frame i.e.

$$E_1(k) = e_w^{T}(k)e_w(k) \qquad k=1,2....k_{max} \tag{5.21}$$

Although the minimization of $E_1(k)$ in the multipulse algorithm produces good quality speech, this is not necessarily an appropriate measure (as will be shown) for variable rate coding since, in this case, the mse from different segments of speech has to be taken into account. Alternative measures can be devised by introducing a form of normalization through division by the quantity

$$S = Y_w^{T} Y_w \tag{5.22}$$

which represents the weighted "energy" of the current speech frame. The SL-SULJ-MUSJ algorithm was used by setting $d_{con} = \infty$, $d_s = 0.0$ and the value of d_L chosen so that the resulting average bit rate/frame was near 7 pulses/frame (i.e. a similar criterion to those of figure 5.29). The pulse pdf using the distortion measure $E_1(k)$ of 5.21 is shown in figure 5.32. Alternative measures used were:

$$E_2(k) = \frac{E_1(k)}{S} \tag{5.23}$$

(figure 5.33) for which $d_L = 0.17$,

$$E_3(k) = \frac{10\log_{10}E_1(k)}{10\log_{10}S} \tag{5.24}$$

(figure 5.34)

$$E_4(k) = 10\log_{10}\frac{E_1(k)}{S} \tag{5.25a}$$

$$E_5(k) = 10\log_{10}E_1(k) \qquad 5.25b$$

(figure 5.35)

$$E_6(k) = (E_1(k))^{1/4} \qquad 5.26$$

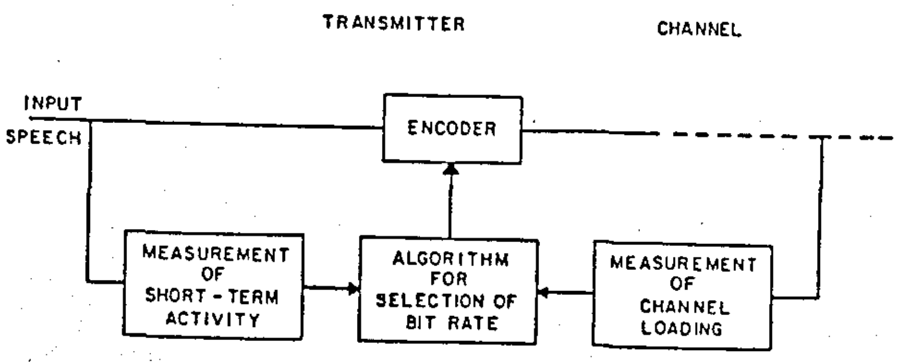
(figure 5.36)

and finally,

$$E_7(k) = \frac{[E_1(k)]^{1/4}}{[S]^{1/4}} \qquad 5.27$$

(figure 5.37)

Note that, since the criterion optimized through the multipulse algorithm (or a monotone function of it) is the same one used in the variable rate algorithm, no local maxima/minima exist for the distortion functions 5.21 and 5.23-5.27. Equation 5.21 represents a weighted m.s.e. criterion and minimization of 5.4 subject to a distortion function d(t,k) given by 5.21, is equivalent to a wmse rate distortion theory minimization result. This scheme also gave the worst subjective performance (this can also be deduced from the corresponding pulse pdf in figure 5.32). The measure of 5.23 assumes that the relative noise power is relevant in pulse allocation. This criterion gave the best subjective performance between these schemes, which was similar to the fixed rate performance. The measure of 5.25 is similar to an SNRSEG measure. Note that under the variable rate algorithm used equations 5.25a and 5.25b are equivalent. The distortion measure of equation 5.27 bears some resemblance to Schroeder's distortion measure (chapters 2 and 4). All measures given by equation 5.24-5.27 produced coded files with subjective quality worse than that of the measure of equation 5.23, with quality deteriorating as the corresponding (pulse) pdf becomes less similar to that of figure 5.33.


TRANSMITTER
CHANNEL
INPUT
SPEECH
ENCODER
MEASUREMENT OF SHORT - TERM ACTIVITY
ALGORITHM FOR SELECTION OF BIT RATE
MEASUREMENT OF CHANNEL LOADING

General block diagram of a variable-rate encoder.

Figure 5.1 [1]

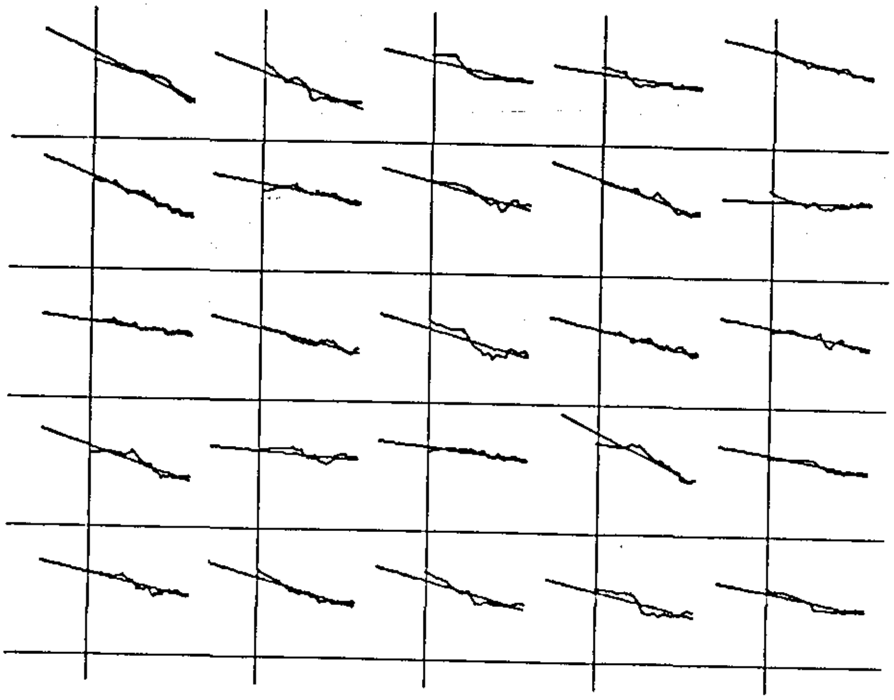


Figure 5.2 Logarithm of distortion (ordinate) as a function of the logarithm of the number of pulses (abscissa) for successive speech frames. Straight lines fitted to the data (via least squares procedures) are also shown. It can be seen that the data, in general, follow a straight line, although local maxima and minima can also be seen.

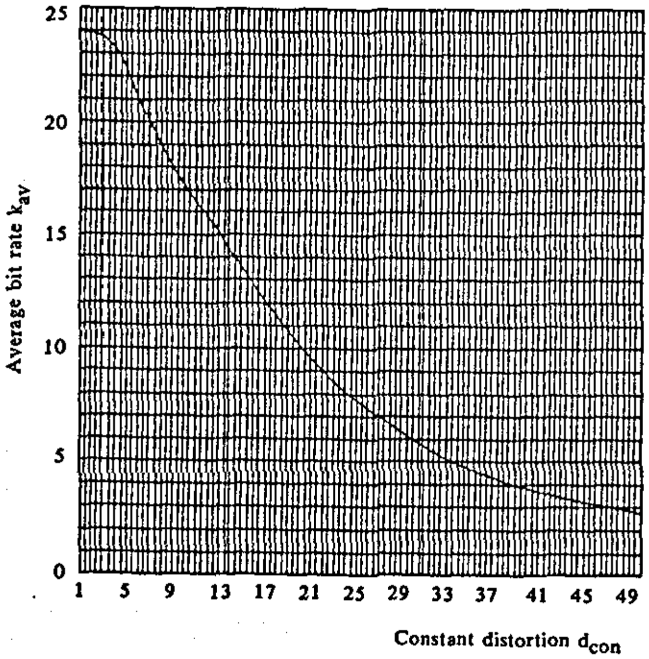


Figure 5.3 Average pulse rate as a function of constant distortion, obtained from training data. This curve is used by the SL algorithm (see text).

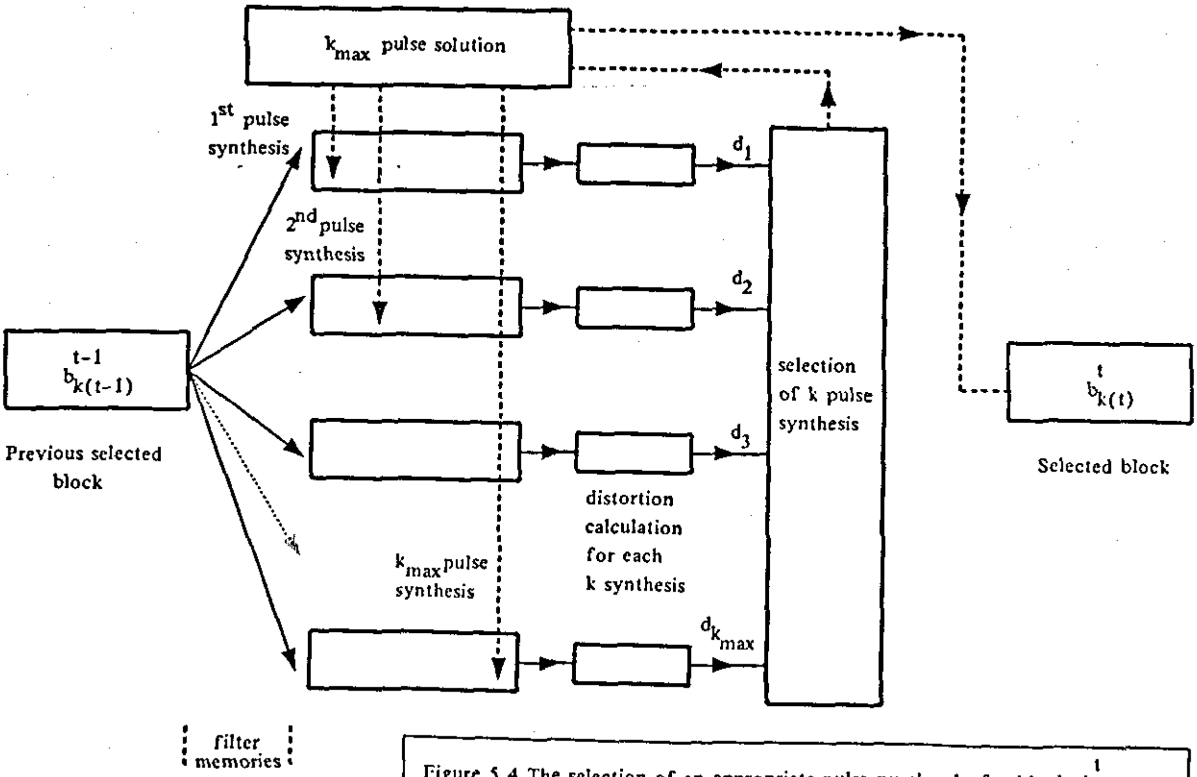


Figure 5.4 The selection of an appropriate pulse number k, for block $b^{t}_{k(t)}$ subject to a predetermined criterion. The k_{max} pulse solution provides solutions for all k less than k_{max}.

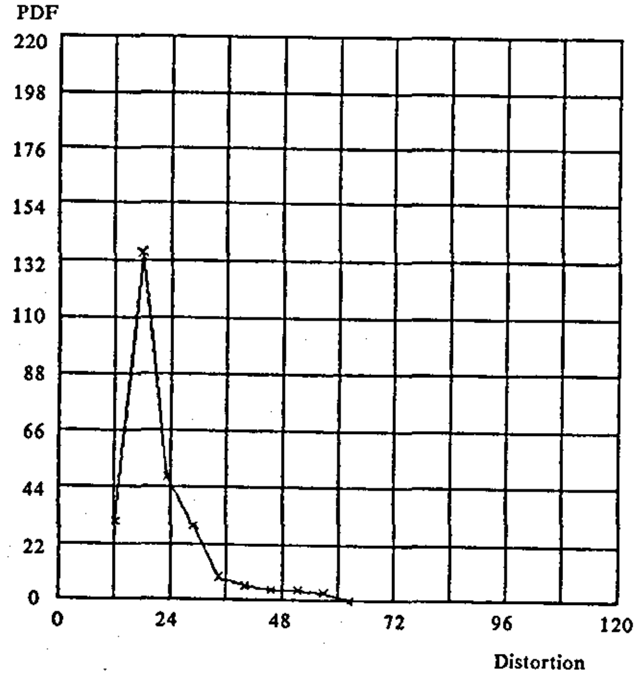


Figure 5.5a Pdf of distortion obtained with the straight line (SL) approximation algorithm. A near constant distortion is obtained.

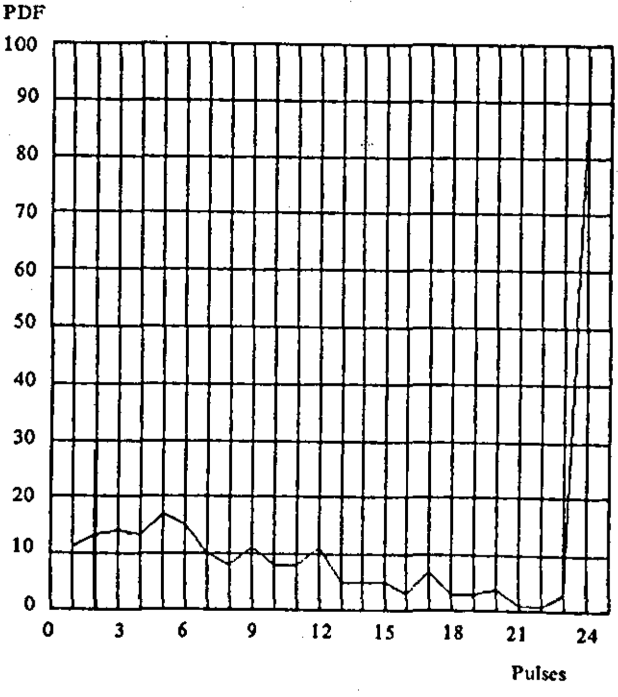


Figure 5.5b Pdf of the pulse rate obtained with the straight line (SL) algorithm. Note the high value at the end of the interval (24 pulses).

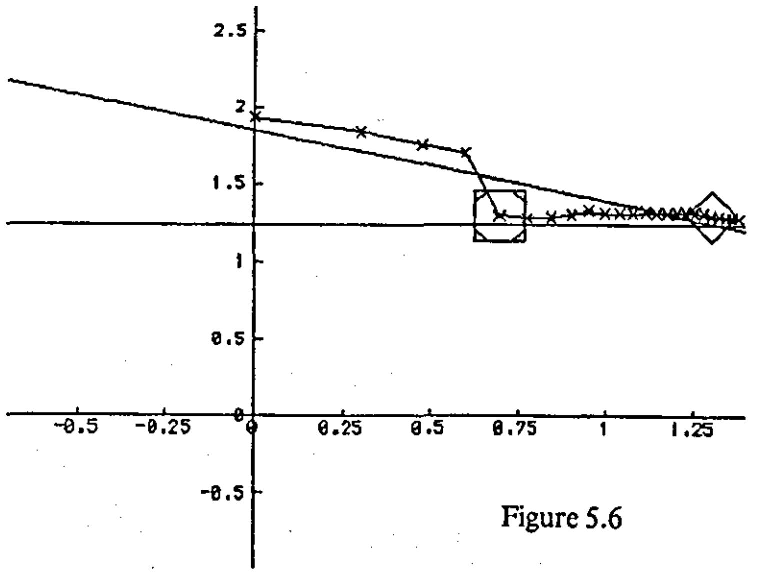


Figure 5.6

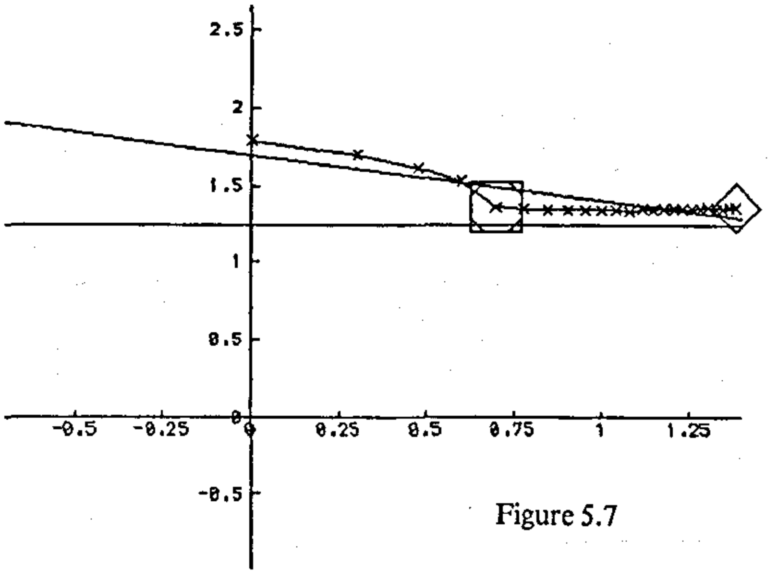


Figure 5.7

Figures 5.6-5.8 Three sample plots of distortion as a function of the bit rate (both on logarithmic scales) for successive speech frames. The constant distortion value, central to the SL algorithm, is shown by the horizontal lines. It can be seen that the constant distortion line intersects the SL line outside the permissible pulse range. The upper pulse constraint is therefore selected (24 pulses, rhombus markers). A maximal decrease in bit rate can be achieved with a minimal increase in distortion if the pulse rates shown by the the octahedron-in-square markers are chosen instead. These new pulse rates were the choises made by the SULJ algorithm.

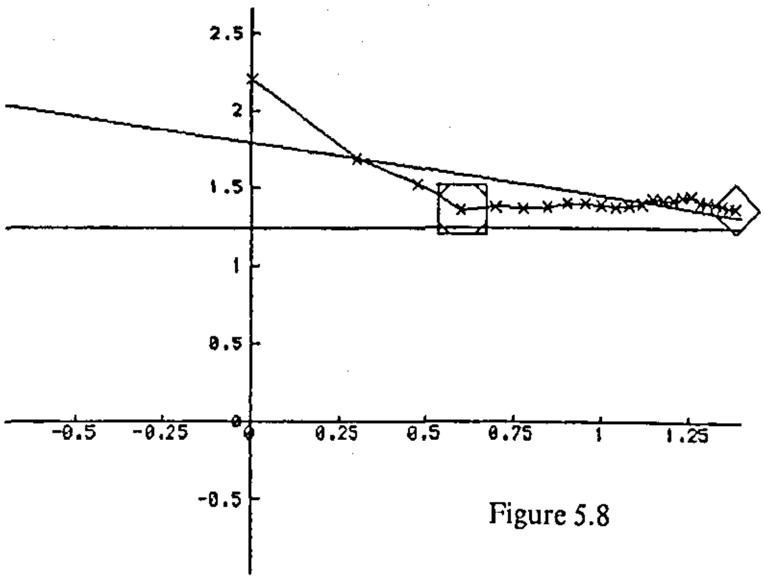


Figure 5.8

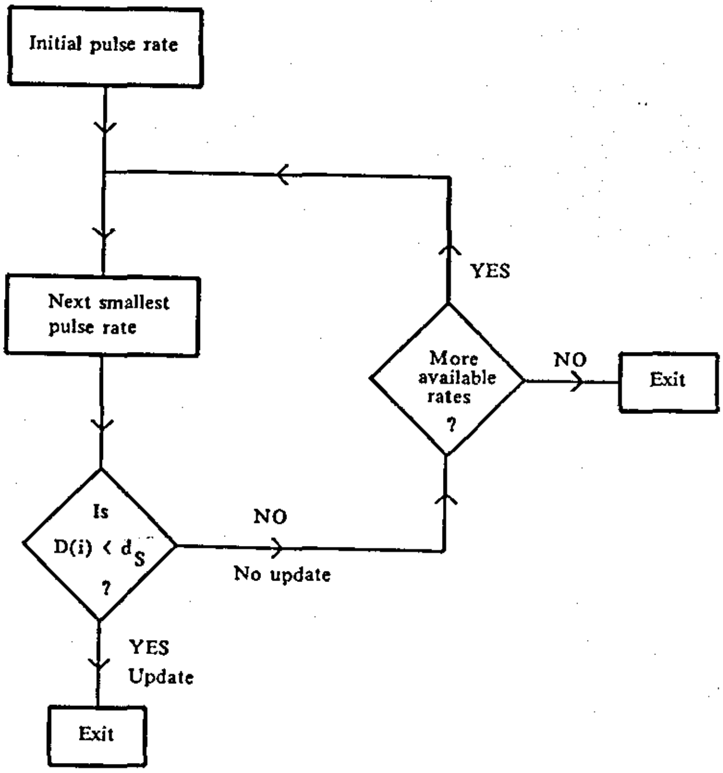


Figure 5.9 SULJ
The single update largest jump algorithm.
(cost is reduced)
D(i) = D(new) - D(old)

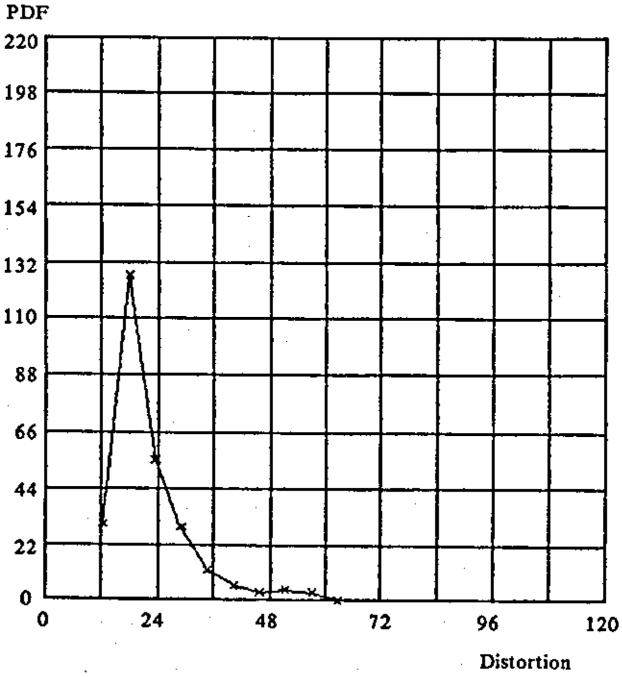


Figure 5.10a Pdf of distortion as a result of applying the SL algorithm followed by the SULJ algorithm. Note the similarity of this distribution with the one in figure 5.5a, employing only the SL algorithm. The threshold parameter (maximum increase in distortion in any one frame, d_s) was set to 2.5.

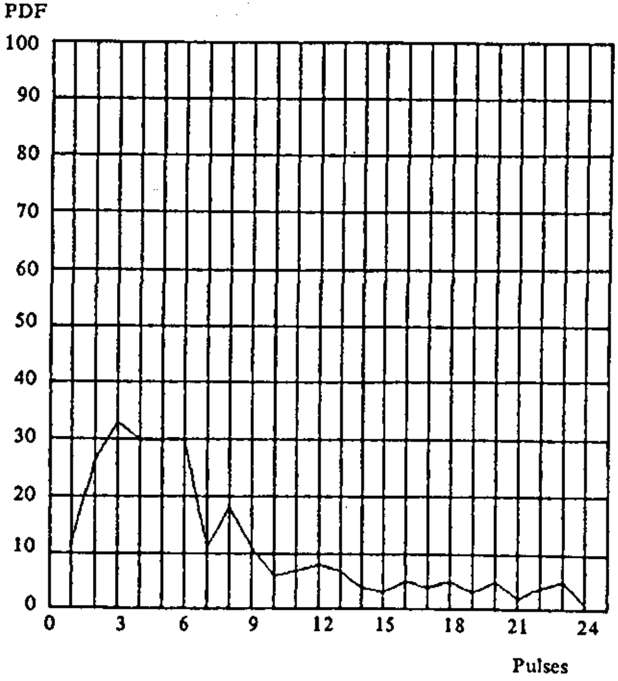


Figure 5.10b *Pulse pdf obtained* with the SL and SULJ algoritms. The average pulse rate is 7.5 pulses per frame. Note that the high pdf value at 24 pulses is no longer obtained.

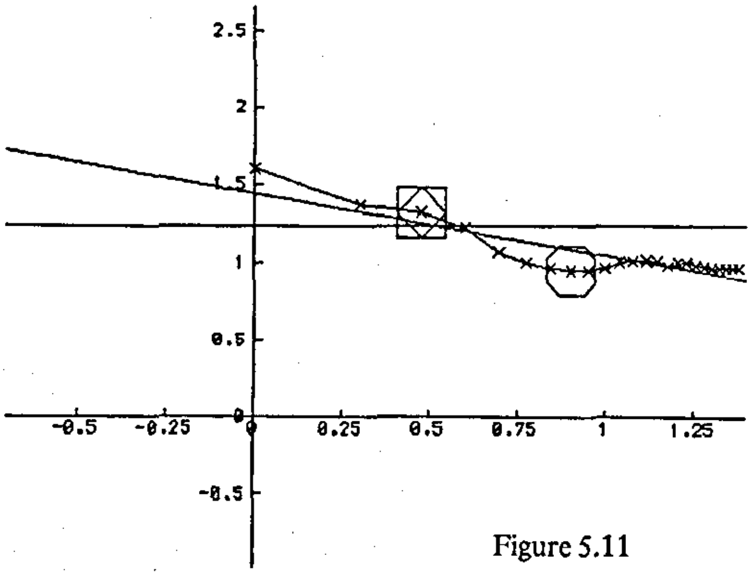


Figure 5.11

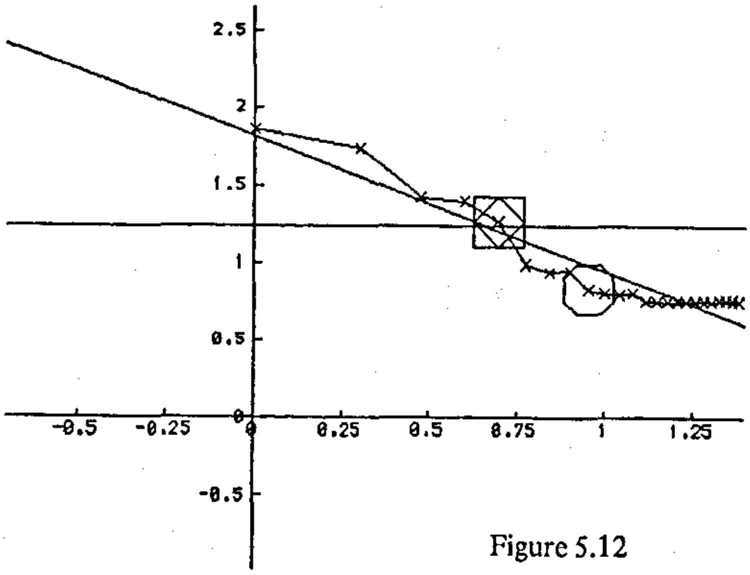


Figure 5.12

Figures 5.11–5.14 Rates selected by the SL algorithm (rhombus-in-square markers) and subsequent corrections made by the MUSJ algorithm (octahedron markers). A mid-range value has been selected by the SL algorithm. A minimal increase in the pulse rate is achieved with a maximal decrease in distortion by employing the MUSJ algorithm.

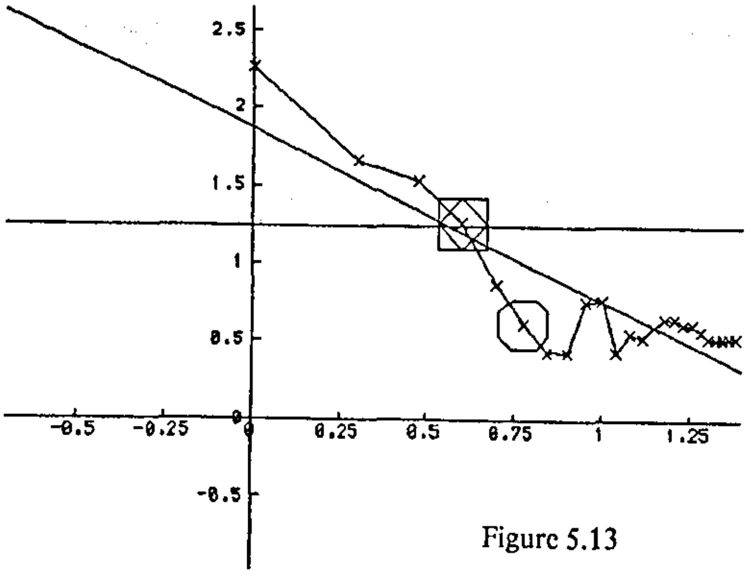


Figure 5.13

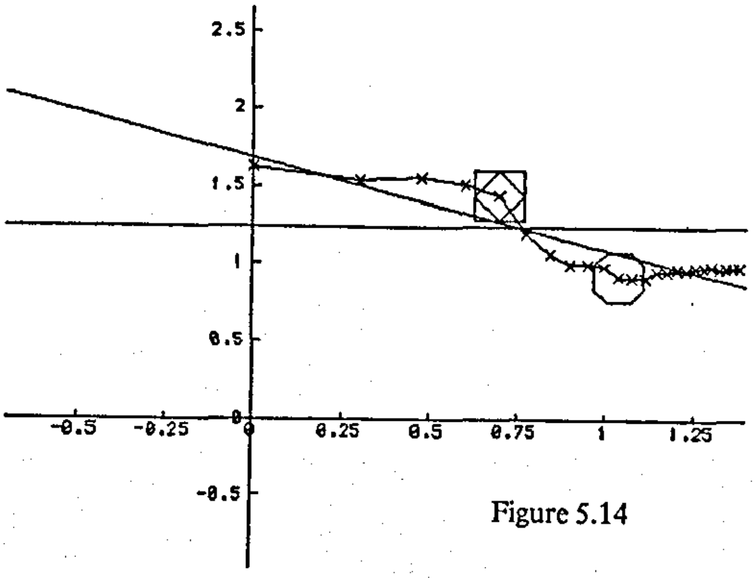


Figure 5.14

For legend see figure 5.12.

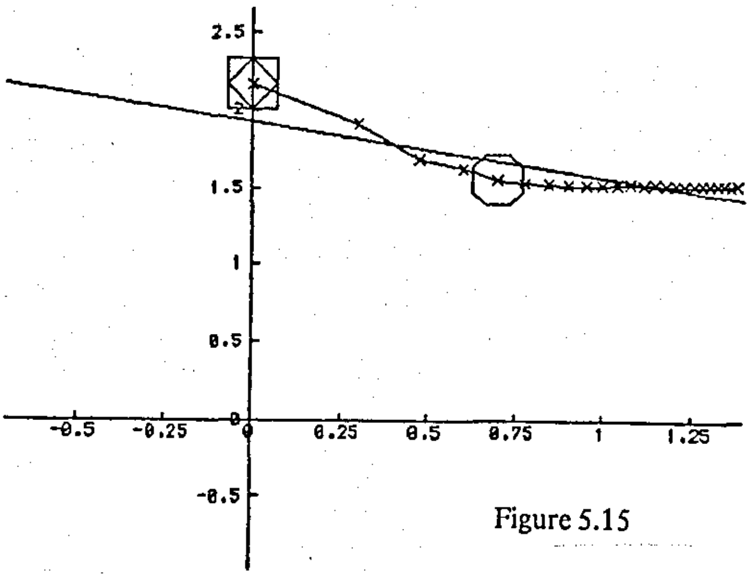


Figure 5.15

Figures 5.15-5.18 As in figures 5.11 -5.14. The selection made by the SL algorithm lies at the lower pulse number constraint (1 pulse allocated). The SL algorithm fails because for a high enough value of preselected distortion, the constant distortion (horizontal) line intersects the SL distortion line outside the permissible pulse range. The MUSJ algorithm corrects for the above effect (octahedron markers).

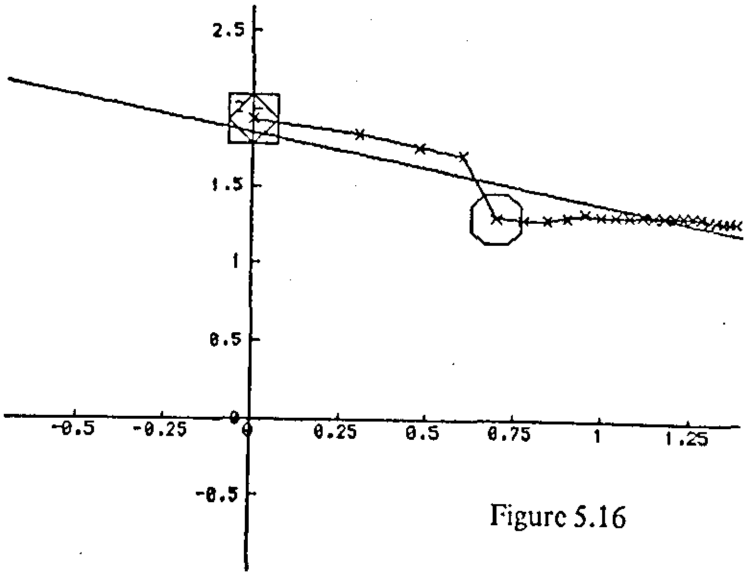


Figure 5.16

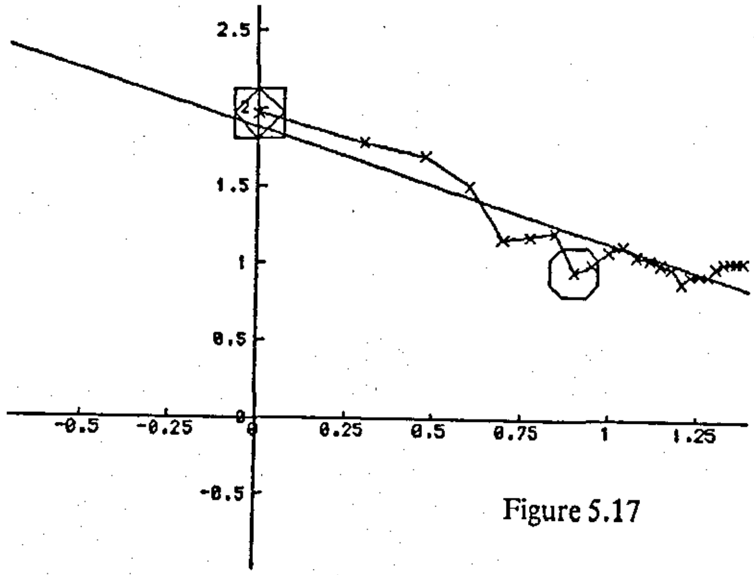


Figure 5.17

For legend see figure 5.15.

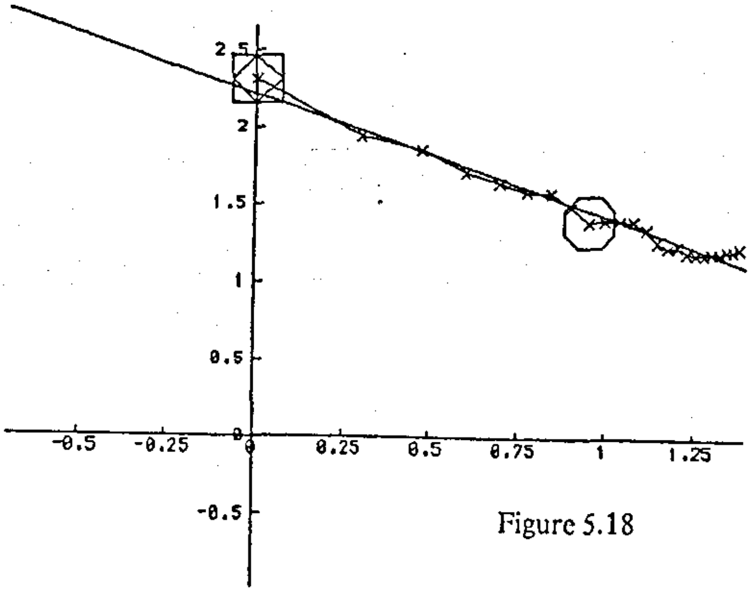


Figure 5.18

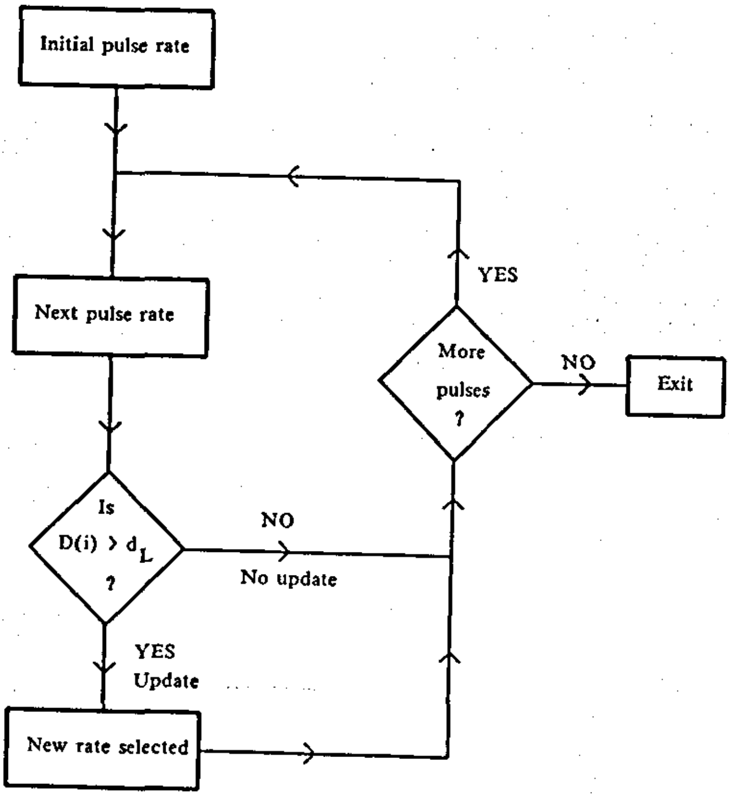


Figure 5.19 MUSJ
The Multiple update
smallest jump algorithm.
(cost is increased)
D(i) = D(old) - D(new)

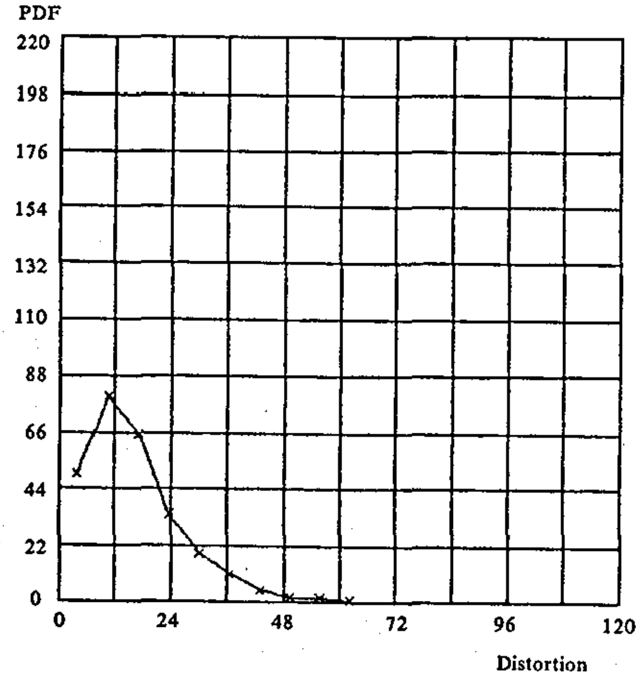


Figure 5.20a Distortion pdf obtained using the SL algorithm, followed by the SULJ and MUSJ algorithms. The distortion is shown to be less than that of figure 5.5a, although a broader range of values is obtained. The SL algorithm was run for a value of d_{con} of 17.5 (as in figures 5.5 and 5.10), a value for d_S of 2.5 and a value for d_L of 2.3. For more details see text.

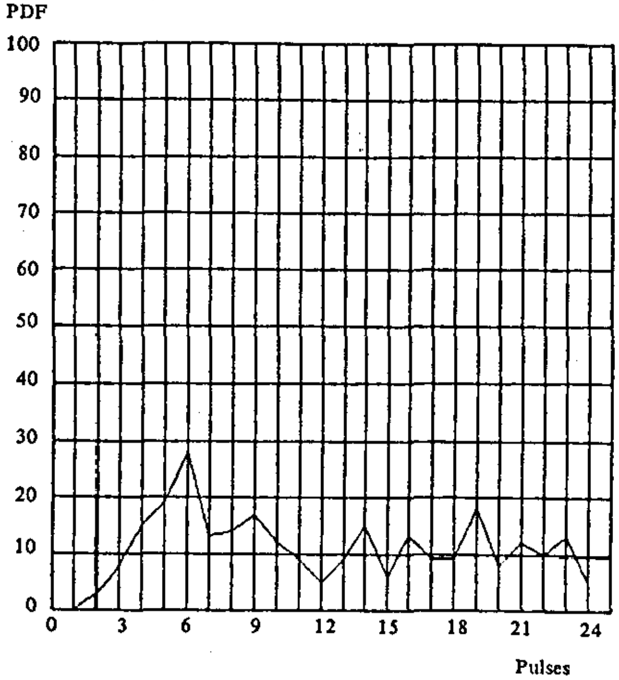


Figure 5.20b Pulse pdf for an average pulse rate near 12 pulses per frame. The combined SL, SULJ and MUSJ algorithms were used. Note the near uniform distribution and the absence of high density values at the ends of the range.

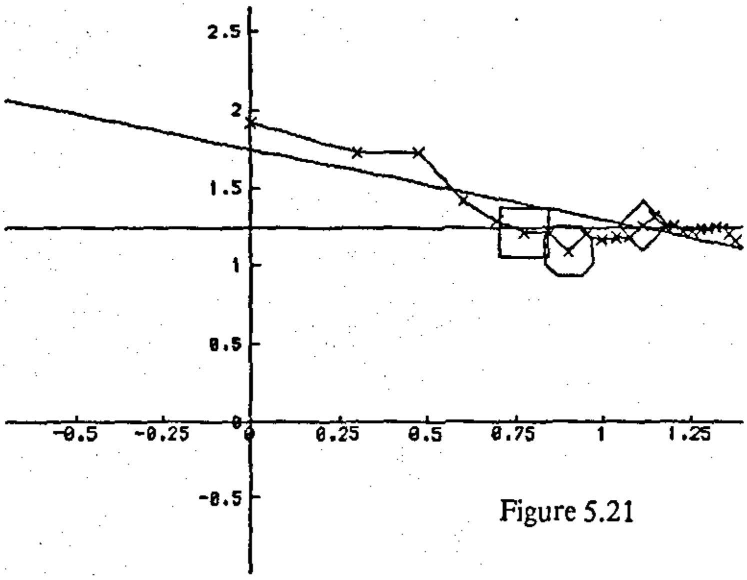


Figure 5.21

For legend see figure 5.23.

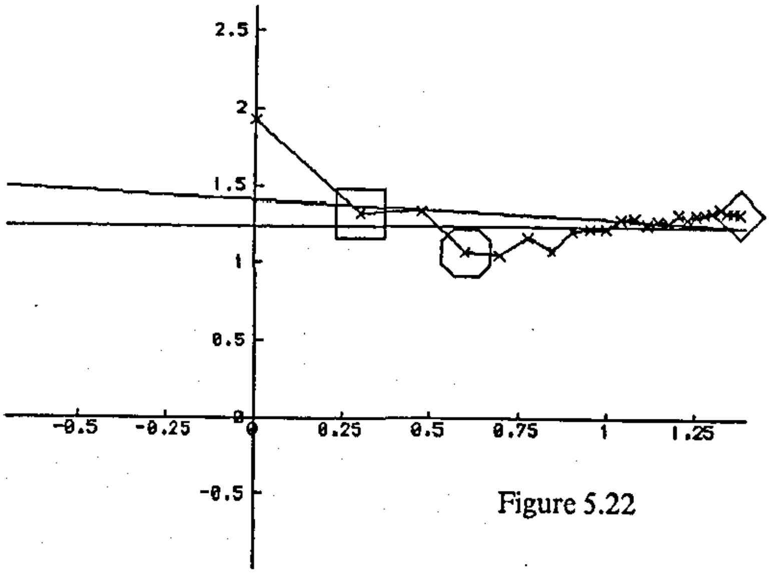


Figure 5.22

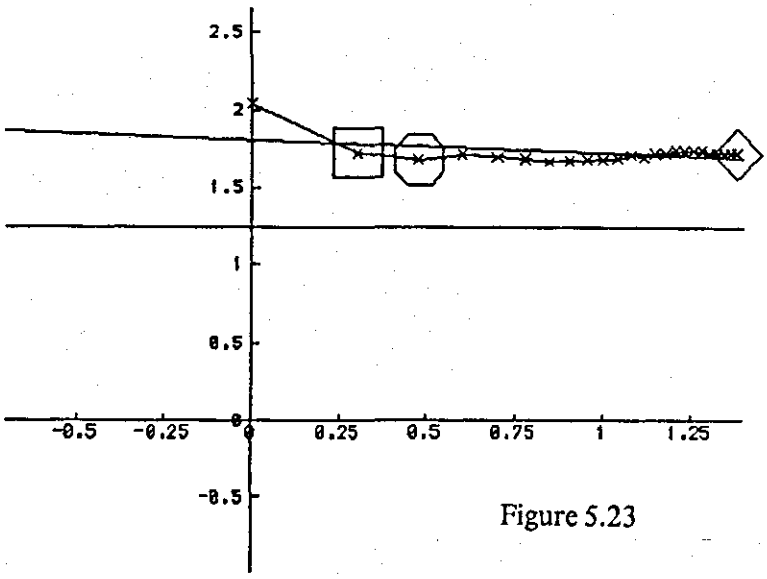


Figure 5.23

Figures 5.21–5.24 The effects of the combined SL–SULJ–MUSJ algorithms are shown, applied in succession.
Rhombus : pulse rate selected by the SL algorithm.
Square : pulse rate selected by the SULJ algorithm.
Octahedron : pulse rate selected by the MUSJ algorithm.

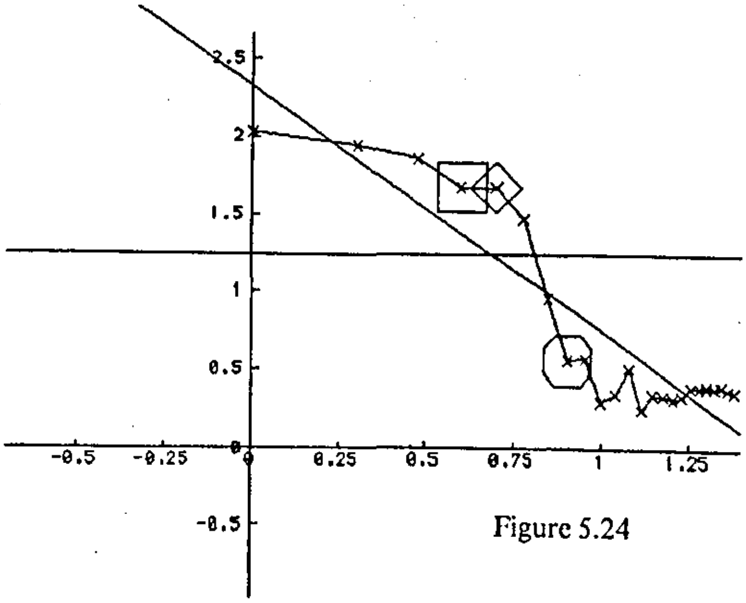


Figure 5.24

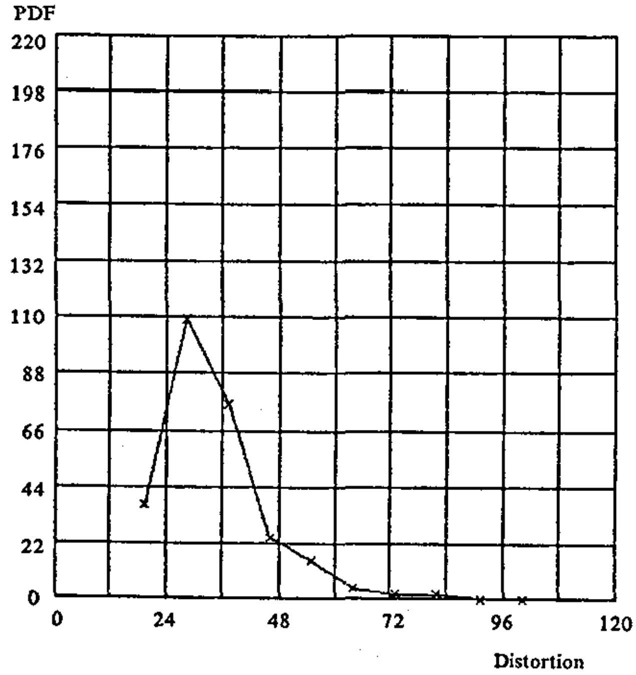


Figure 5.25a As figure 5.5a but with a target distortion of 30.0 instead of 17.5. The resulting average pulse rate is 6.9 pulses per frame as predicted from the plot in figure 5.3.

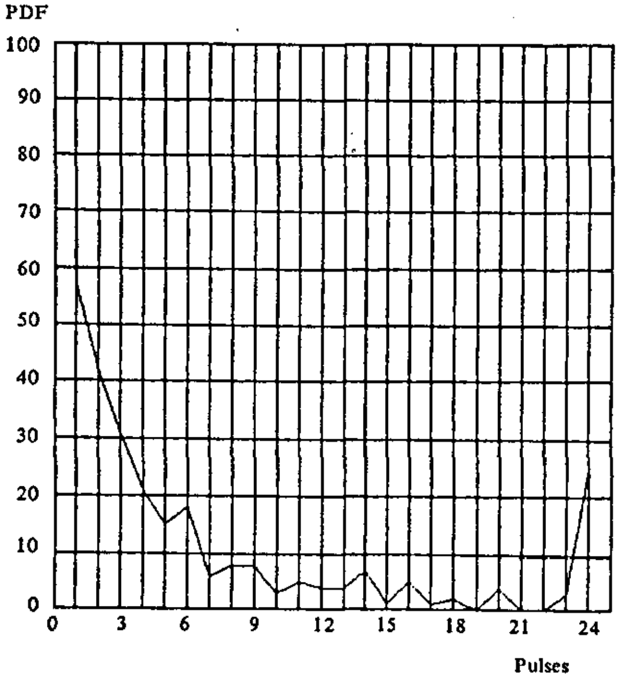


Figure 5.25b Resulting pulse pdf from using the SL algorithm with a target average pulse rate of 7 pulses per frame and an average "constant" distortion of 30.0. Note the large density value at the lower end of the pdf.

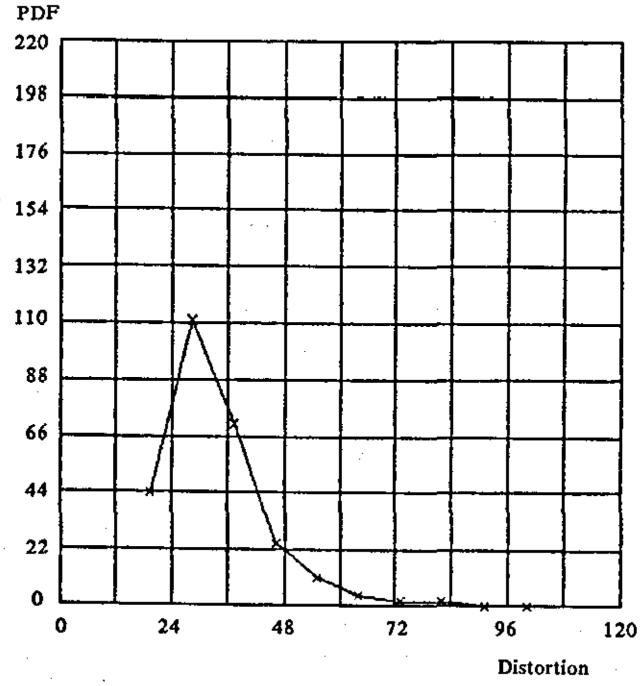


Figure 5.26a Distortion pdf obtained using both the SL algorithm and the SULJ algorithm. A threshold value of 0.0 was used in the SULJ algorithm.

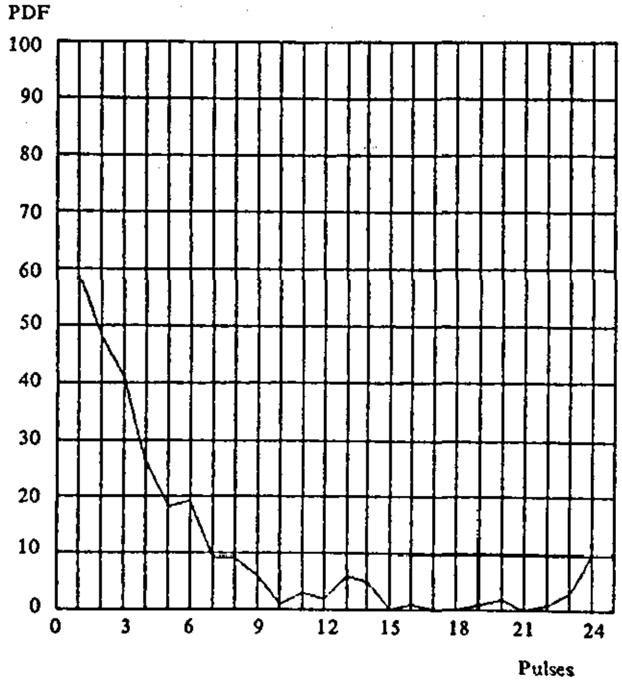


Figure 5.26b Pulse pdf. The avrage pulse rate was reduced to 5.2 pulses per frame as opposed to 6.9 pulses per frame of figure 5.25, with zero increase in distortion. This is due to the nonmonotonicity of the individual distortion versus pulse rate curves.

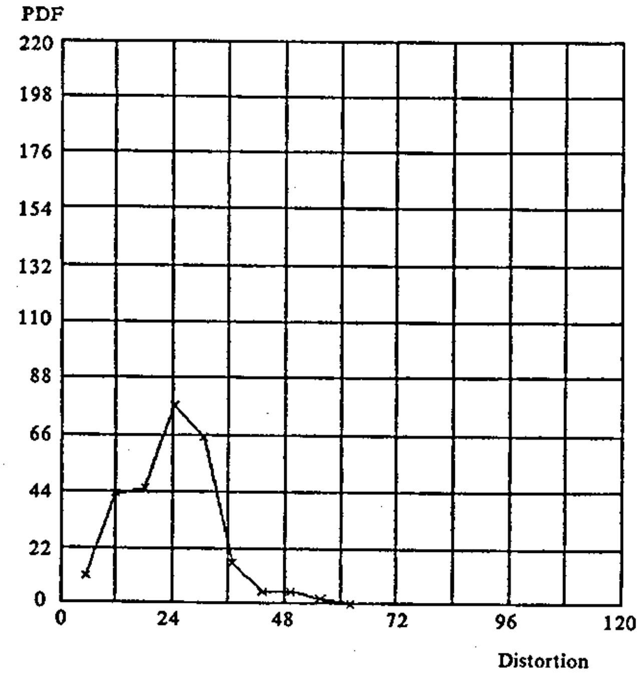


Figure 5.27a The full SL, SULJ, MUSJ algorithm was applied in this case. The various algorithm parameters were: d_{con} =30.0, d_S =0.0 and d_L =23.0. Note that the overall distortion is higher than the equivalent one from the fixed rate algorithm (figure 5.28).

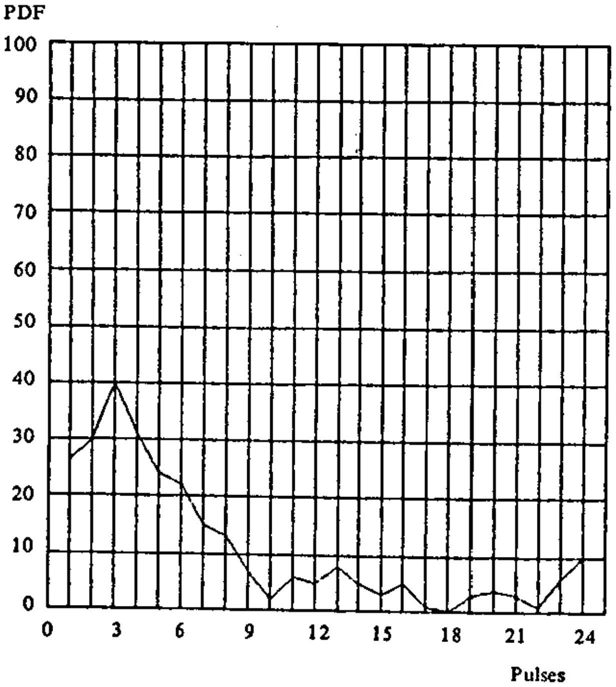


Figure 5.27b. The pulse pdf for the combined algorithm. Parameters as in figure 5.27a. Note the high density values at the lower end of the pdf.

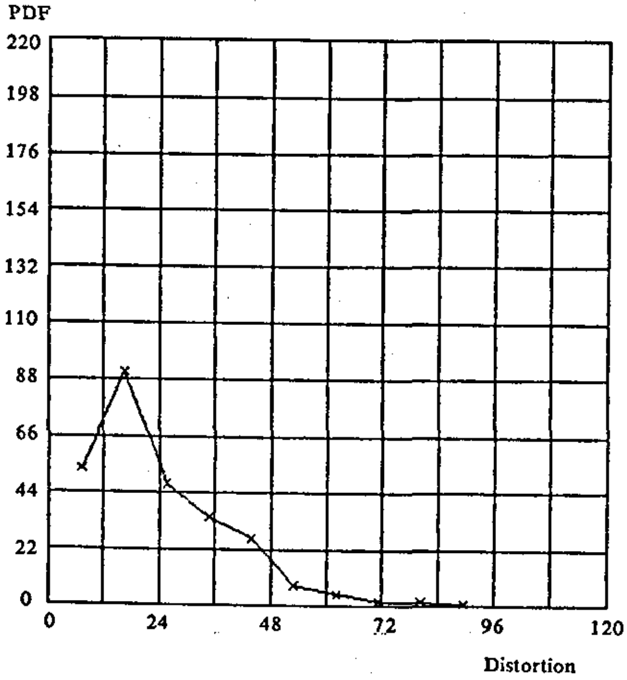


Figure 5.28 Distortion pdf for the fixed rate configuration.(7 pulses per frame).

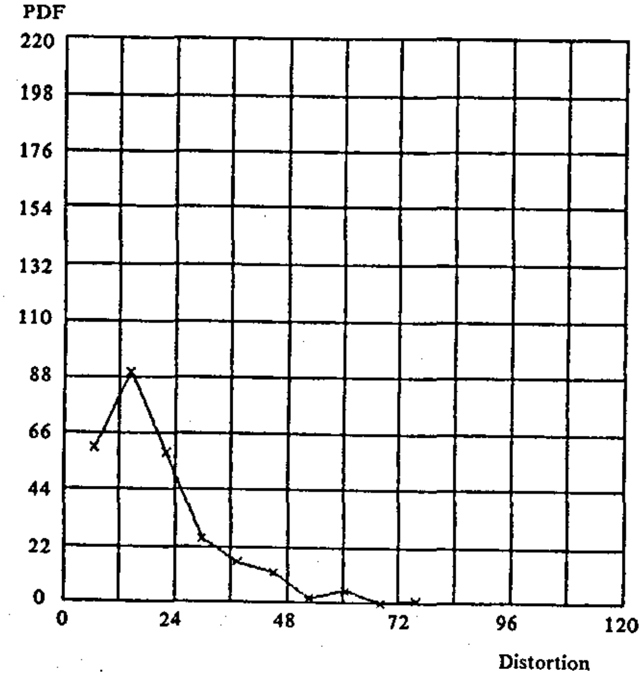


Figure 5.29a Reduced distortion obtained by constraining the maximum number of pulses to 12 (so that the target value, 7, is at the middle of the range). Algorithm parameters can be found in the text. Compare with figures 5.27a and 5.28.

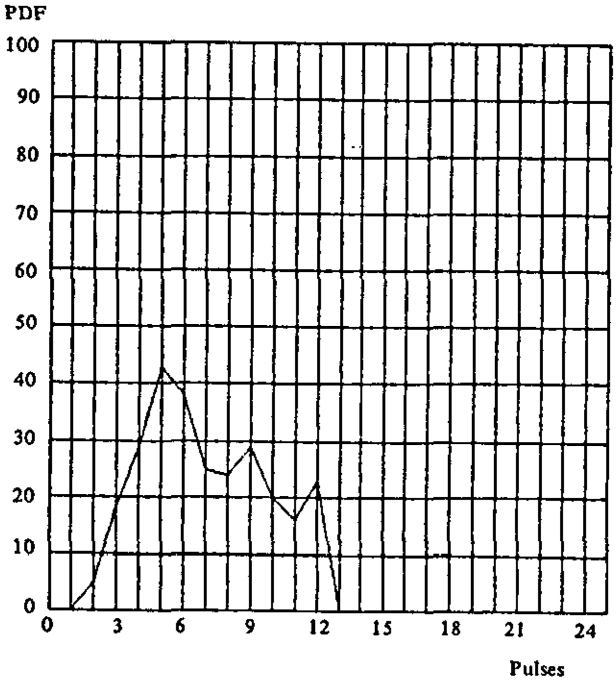


Figure 5.29b Pulse pdf for an average pulse rate of 7 pulses per frame. The maximum allowable rate is 12 pulses.

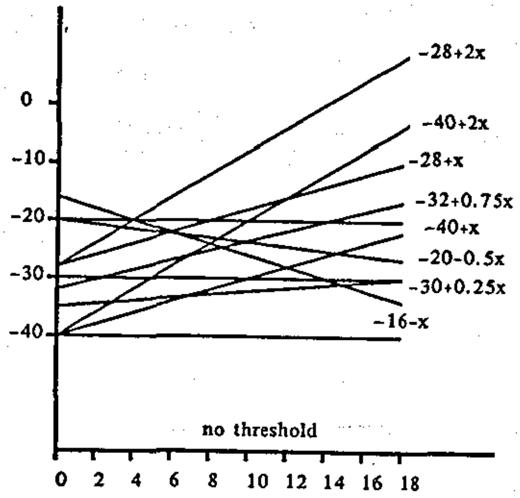


Figure 5.30 Threshold functions employed in the distortion measure to select an appropriate pulse distribution in time. The best function was determined through subjective tests. The best performance was obtained with the thrreshold set at (-32+0.75x) although some of the other functions shown such as (-28+x), (-30), (-35+0.25) gave a similar performance. The worst performance was obtained with the threshold at (-16-x).

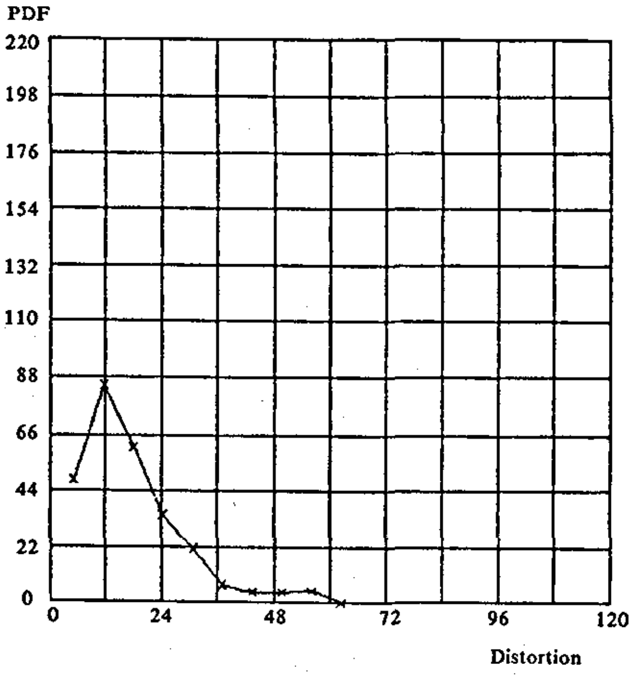


Figure 5.31a Distortion pdf obtained using a "non-real time" algorithm.

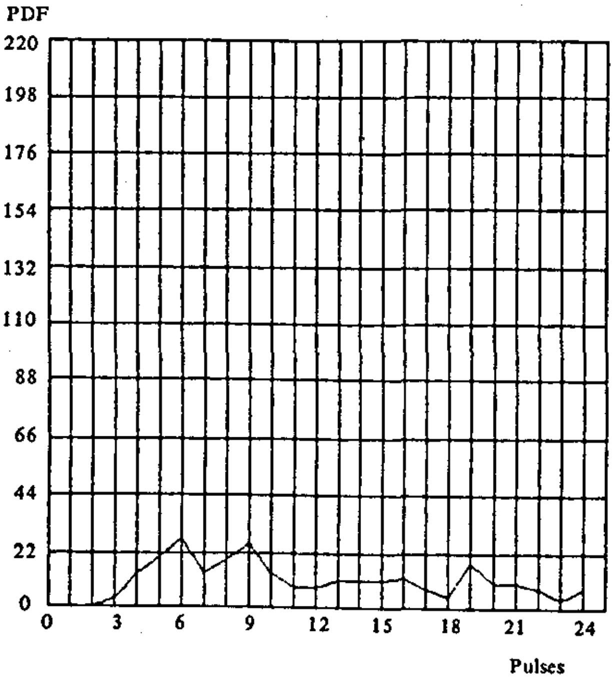


Figure 5.31b Pulse pdf obtained using a "non-real time" algorithm. Note the similarity with figure 5.20b.

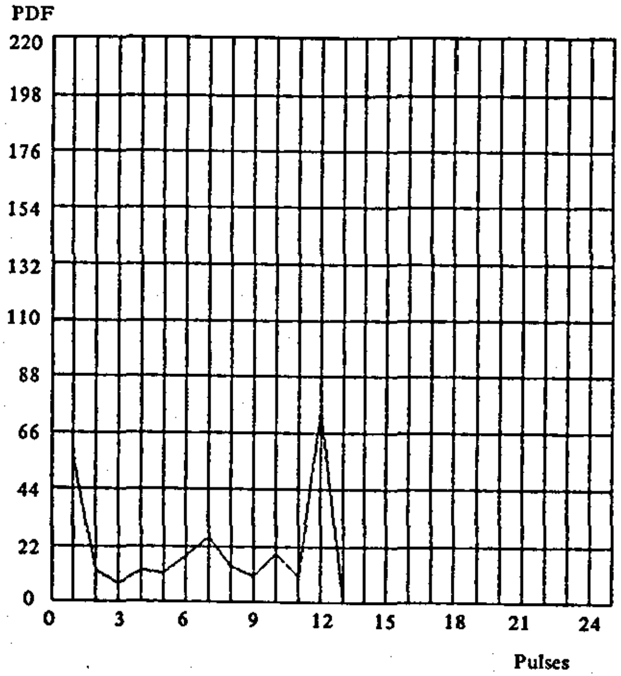


Figure 5.32 A variable rate algorithm employing a wmse criterion. For algorithm parameters see text. Note the concentration of high density values at the end of the range, signifying a high overall distortion and a bad rate selection model.

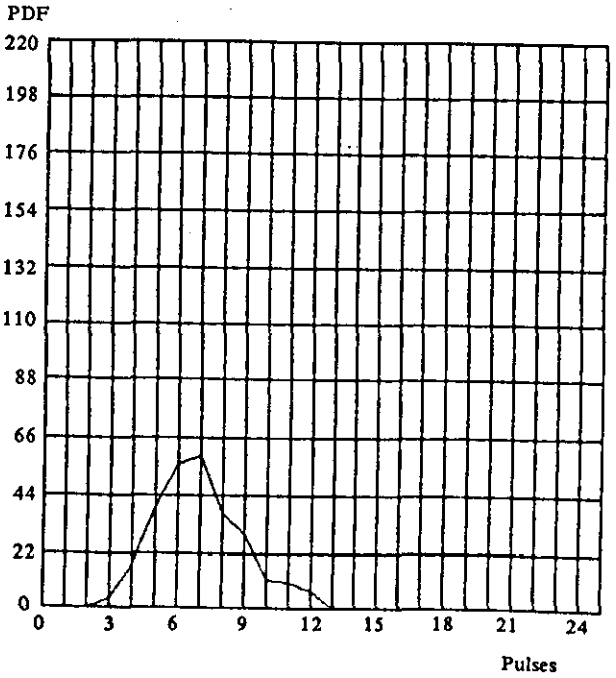


Figure 5.33 The pulse distribution with the best subjective performance among the algorithms based on a wmse criterion.

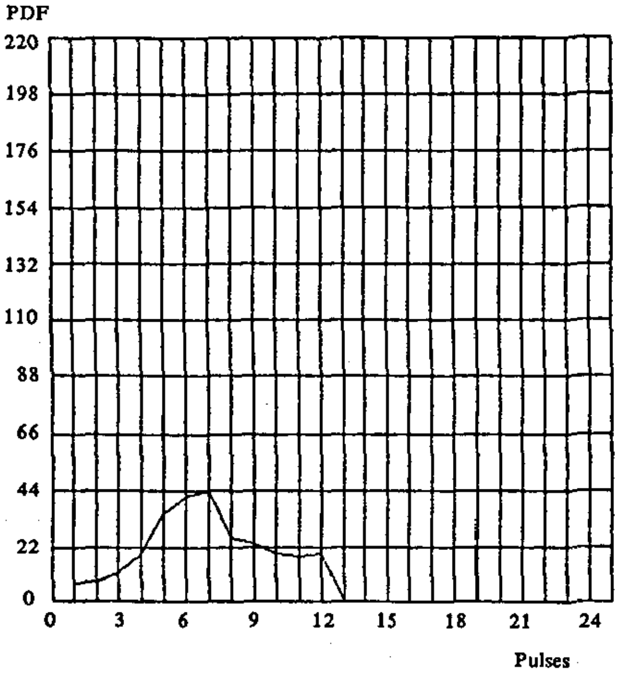


Figure 5.34 Pulse pdf obtained with a variable rate algorithm based upon a wmse and logarithmic compression.

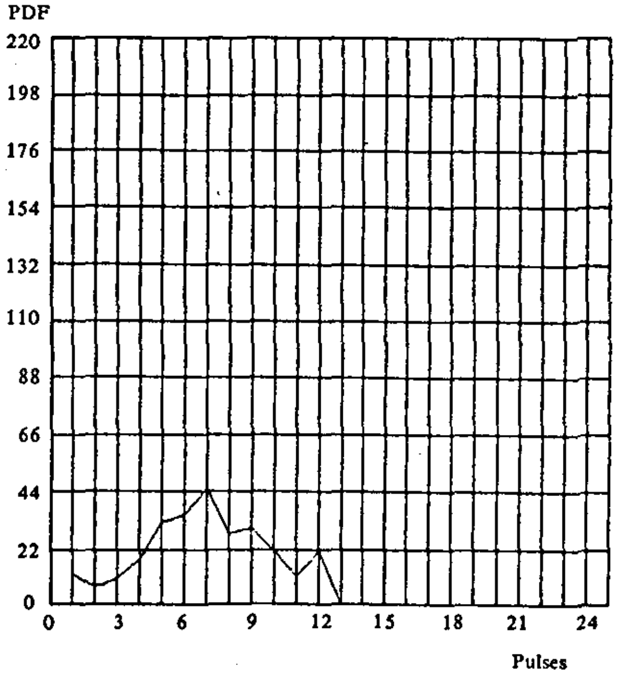


Figure 5.35 Pdf of pulse rate obtained with a variable rate coder employing a (weighted) SNRSEG measure.

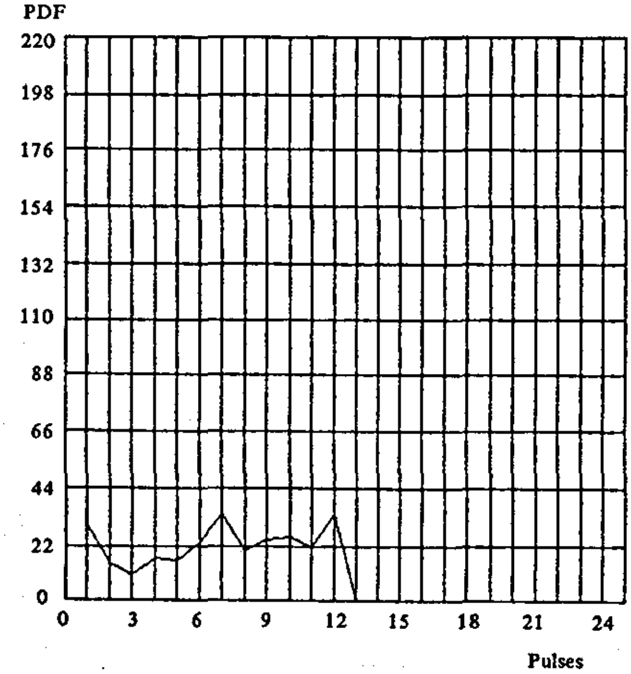


Figure 5.36 Pdf of pulse rate obtained with a wmse criterion and "neural" compression.

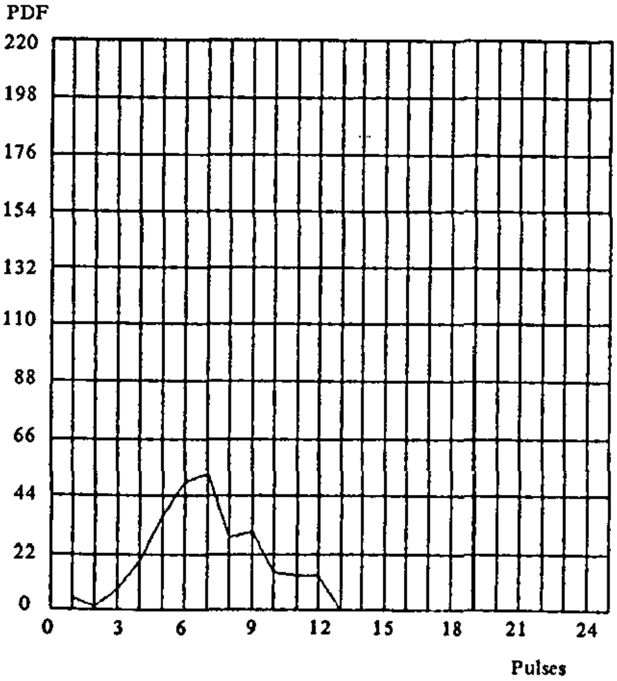


Figure 5.37 Pulse Pdf with a distortion criterion employing a normalized loudness measure (see text).

1. N.S. Jayant "Variable Rate Speech Coding: A review" International Conference on Communications - ICC 84, May 1984, Vol. 3, pp1614-20.

2. J.J. Dubnowski, R.E. Crochiere "Variable Rate Coding" IEEE Proc. ICASSP 1979 pp445-448.

3. R.V. Cox, R.E. Crochiere "Multiple User Variable Rate Coding for TASI and Packet Transmission Systems" IEEE Trans on Commun. Vol. COM-28, No. 3, March 1980.

4. Y. Yatsuzuca "Highly Sensitive Speech Detector and High-Speed Voiceband Data Discriminator in DSI-ADPCM Systems" IEEE Trans on Comms, Vol. COM-30, No. 4, April, 1982, pp739-750.

5. Y. Yatsuzuka "High Gain Digital Speech Interpolation with Adaptive Differential PCM Encoding" IEEE Trans. on Comms. Vol. COM-30, No 4, April 1982 pp750-761.

6. J.B. O'Neal Jr., R.W. Stroh, "Differential PCM for Speech and Data Signals" IEEE Trans on commun Vol. COM-20, No 5, Oct. 1972 pp900-912.

7. J. Makhoul, M. Berouti, M. Krasner "Time and Frequency Domain Noise Shaping in Speech coding" IEEE Proc. ICASSP 1981 pp611-614.

8. E.D. Frangoulis, K. Yoshida, L.F. Turner "Adaptive differential pulse-code modulatiorn with adaptive bit allocation" IEE Proc. Vol. 131, pt. F, No 5, Aug, 1984 pp542-548.

9. M. Honda., F. Itakura "Bit Allocation in Time and Frequency Domains for Predictive Coding of Speech" IEEE trans on Acoust. Speech and Signal Proc. Vol-ASSP-32 No 3, June 1984, pp465-473.

10. T. Bially, B. Gold, S. Seneff "A Technique for Adaptive Voice Flow Control in Integrated Packet Networks" IEEE Trans on Commun. Vol. COM-28 No 3, March 1980, pp325-333.

11. G.G. Langenbucher "Efficient Coding and Speech Interpolation: Principles and Performance Characterization" IEEE Trans. on Commun. Vol. COM-30, No. 4, April 1982, pp769-779.

12. H.L. Gerhäuser "Digital Speech Interpolation with Predicted Wordlength Assignment ((PWA). IEEE Trans on Commun. Vol. COM-30, No. 4, April 1982 pp762-769.

13. J.G. Gruber "A Comparison of Measured and Calculated speech Temporal Parameters Relevant to Speech Activity Detection" IEEE Trans on Commun. Vol. COM-30, No. 4, April, 1982, pp728-737.

14. N.S. Jayant "Variable Rate ADPCM Based on Explicit Noise Coding" Bell System Technical Journal Vol. 62, No. 3, March 1983 pp657-677.

15. G.Rebolledo, R.M. Gray, J.P. Burg "A multirate Voice Digitizer based upon Vector Quantization" IEEE Trans on Commun. Vol. COM-30, No. 4, April, 1982, pp721-727.

16. T. Araseki, K. Ozawa, S. Ono, K. Ochiai "Multipulse excited speech coder based on maximum crosscorrelation search algorithm" IEEE proc. Global Telecom. Conf. 1983, pp794-798.

17. P. Kroon "Time Domain Coding of (NEAR) Toll Quality Speech at rates below 16 kb/s" Ph.D. report, Delft Univresity of Technology, Makelweg 4, 2628 CD Delft, The Netherlands, March, 1985.

CHAPTER 6:

A NEW APPROACH TO LOW BIT RATE SPEECH CODING

6.1 INTRODUCTION

In recent years a variety of coders such as multipulse LPC have been designed to bring the bit rate of communications quality speech below the 9.6 kb/sec barrier of conventional waveform coders. These coders combine some vocoder features into the design thus allowing a hybrid combination of waveform coding and vocoding. These particular voice coding techniques take advantage of the spectral envelope and pitch related redundancies in the signal. The waveform coding and vocoding techniques are integrated together through an analysis by synthesis process whereby, a speech model is chosen, whose parameters are tuned through the analysis by synthesis error minimization, to yield a waveform "close" to the input signal. Thus, although a vocoder model is used, a close approximation to the waveform is recovered at the output of the decoder. All such coders are designed around predictive coding and, as a consequence, suffer from the effects of quantization noise feedback and poor noise-shape control over the spectrum, at these low bit rates.

For the higher bit rates that produce toll quality speech, subband coding has proved beneficial in reducing the effect of the above degradations in the speech signal, by splitting the prediction burden between frequency and time domain. A major advantage of coding the subbands is the versatility and choice provided, for coding the subband signals: Different bit rates can be used for each band and also different coding techniques i.e. different coders can be employed for each band in order to minimise and mask the distortion which is now confined to each subband.

The introduction of subband coding is not without disadvantages though: By splitting the signal into subbands some problems particular to this form of coding arize. Amongst others these are: An increase in the side information, increased redundancy by specifying the same speech information in each band separately (e.g. the pitch structure), "empty bands" i.e. bands for which very little information can be transmitted for reconstruction (e.g. only the power level) etc. These problems do not necessarily make subband coding a poor choice as, in most cases, they can be alleviated with careful design and a little increase in complexity. For example, the increase in side information can be absorbed into the bits used for time domain prediction since prediction gain and subband gain exploit the same characteristics of the speech signal, namely the nonflatness of the speech spectrum: Lower bit rates need be used for the side information of the time domain prediction when frequency domain prediction is also used.

In the work to be presented, the subband approach was used to split the speech signal into eight contiguous bands using a tree-structured QMF bank, although the eighth band of 3500-4000 Hz was always discarded. These signals were then coded independently or in combination with each other, so that the maximum perceptual improvement could be obtained. The coding techniques used were analysis by synthesis techniques centered around the multipulse algorithm. (Appendix F). For these algorithms a set of LPC coefficients needs to be calculated for each band. The LPC coefficients form one part of the side information. The other part of the side information, which is necessary for the operation of the algorithms, is related to the short term variance of the subband signals: considering the low bit rates for which the coder is intended (4.8-9.6 kbit/sec) it is necessary to apply an adaptive bit allocation strategy. The set of parameters to be used must be able to predict the noise level (m.s.e.) in each band. It is also customary to quantize the pulse amplitudes in multipulse algorithms using an AQF quantizer. It is desirable that the side information necessary to predict the m.s.e. in each band can also be used for normalization purposes in the AQF algorithm.

6.2 SIDE INFORMATION: LPC FILTERS

With the intention of using Vector Quantization for the encoding of the LPC parameters, 4th order filters derived using the Burg method were employed in each band. This could be considered excessive in view of the fact that it leads to a total of 28 LPC coefficients to encode the spectrum. It is normally considered that a 12th order filter (12 coefficients) is adequate to remove the short term (spectral envelope) redundancy from the full-band speech signal. In addition, the time domain prediction is supplemented by the frequency domain redundancy removal inherent in the subband structure. It was observed though that the use of smaller order filters in the higher bands resulted in rather short impulse responses of the filters, which combined with a small number of pulses, gave rise to gaps in the recovered waveforms in those bands and a reverberant quality in the decoded speech signal. What is important from a transmission point of view is the total number of bits allocated for the envelope information in the speech signal, and not the total number of parameters. Vector quantization (VQ) was therefore applied to encode a large number of parameters with a rather small number of bits. Note that due to the averaging nature of VQ during the training phase, and for a small size codebook, the corresponding subband spectrum will not be as finely reproduced as a 4th order analysis implies but would perhaps be equivalent to a smaller order filter, arising from a "smoother" envelope.

There are several ways to encode the speech envelope information in the subbands. Three different methods are described below.

The first method, which although not implemented is worth mentioning here, combines together all the side information necessary, namely the short term variances per band as well as the short term envelopes per band. This is as follows:

First, a 12th order LPC analysis is performed over the fullband signal, taking into account the appropriate overall delay introduced by the QMF structure so that the block of speech represented by this full-band envelope coincides with the blocks of subband signals to be encoded. The 12th order filter can be used to derive the envelope of the speech signal in the frequency domain:

The LPC model assumes that

$$S(Z) = \frac{GU(Z)}{A(Z)} \qquad 6.1$$

where S(Z) is the speech signal, U(Z) the excitation signal (either a unit impulse or white noise of unit variance) and G/A(Z) is the speech model. From the above equation:

$$|S(f)|^2 = \frac{G^2}{|A(f)|^2} \qquad 6.2$$

G^2 can be given by:

$$G^2 = R(0) - \sum_{k=1}^{P} a_k R(k) \qquad 6.3$$

were the {R(k)} are the autocorrelation coefficients of the speech signal or, alternatively, those of the impulse response of 1/A(Z). The autocorrelation coefficients can be derived directly from the speech signal or, to avoid windowing the speech sequence, they can be derived form eq. 3.2-60 given again below:

$$R(i) = \sum_{k=1}^{P} a_k R(|i-k|) \qquad 1 \leq i < P \qquad 6.4$$

to within a constant factor. The factor can be estimated by forcing R(0) to obey

$$R(0) = E^2 \qquad 6.5$$

where E^2 is the speech power in the block. This is not necessary though, since, in this case, only the relative powers in the subbands are required. To extract the relative power in a particular subband and also the short term envelope in that band, the fourier transform representation can be multiplied by the amplitude response of the appropriate bandpass operation of the QMF structure in that frequency region. (For simplicity, rectangular passband filters can be assumed if desired). Let B(f) be the equivalent QMF response, such that:

$$SB(f) = S(f)B(f) \qquad 6.6$$

where SB(f) is the subband signal in question and S(f) the fourier representation of the fullband signal. From 6.2, the (relative) power per subband is given by:

$$\text{relative power/band} = \sum_{F} \frac{B^2(f)G^2}{|A(f)|^2} \qquad 6.7$$

where F is the region over which $|B(f)|^2$ is large enough to give a non-negligible contribution to the summation.

Due to the fact that the LPC envelope is a better fit on the formants than on the valleys, the estimate in the valleys will probably be a bit higher than the true value but this could be taken care of in the bit allocation procedure.

To obtain the appropriate filter for the subband,the bandpass envelope itself can be shifted to the baseband, taking into account of any frequency inversion effects that would have been introduced by the QMF structure. The envelope is then uniformly expanded to cover the entire 0-$f_s/2$ range where f_s is the cutoff frequency (figure 6.1) and then appropriate autocorrelation coefficients can be calculated by inverse fourier transform of the speech bandpass envelope. From these, using the autocorrelation method of LPC analysis, appropriate LPC coefficients of any desired order can be calculated for that band. The Burg or (stabilized) covariance method can be used to obtain the initial full band signal envelope, so that the final estimate will not suffer from the inaccuracies of the autocorrelation method, arising from block end effects in the time domain.

In order to quantize the subband signals (or obtain a regenerated version at the receiver) the true power level rather than a scaled version of it is needed, and therefore the total power in the full-band block under consideration has to be derived, quantized and transmitted. From this estimate and the relative powers/band, the true power per subband (here the r.m.s. value) can be derived.

From the above discussion it follows that in all, the side information required is identical to the one used to encode the full-band signal in conventional multipulse coding.

The second method is similar to the above in the sense that the subband filters are derived form a wider band representation of the signal. In this method though the subband filters are calculated directly from the subbands: Appropriately shifted blocks of fullband speech are used to derive a 12th order filter as in conventional LPC coding. In addition, associated with this full band filter seven (note that the 8th band of 3500-4000 Hz is always discarded) 4th order filters are derived, one from each of the corresponding blocks of subband signals. The 12th order filter and the subband filters are time aligned to describe the same blocks of speech.

The full-band filter is then vector encoded using the LBG algorithm, with one modification: as well as the final centroids, the indices of the members of each cell are also stored, i.e. each codeword is associated with those members of the training set for which that codeword represents the optimal coded value. From before, for each member of the full band filter set, 7 subband filters have been associated. Therefore, the above subdivision of the full-band filter set into dirichlet regions also partitions each set of bandpass filters into an equal number of "optimal" regions. For each of the sets of subband filters the centroid of each region is calculated and this represents the subband codeword associated with the members of the partition. Note that this algorithm produces 7 subband codewords for each full band codeword. When a particular full-band codeword is selected, the 7 associated subband codewords are automatically selected. This completes the training session. To encode the filters, first, the corresponding full-band filter is derived and the optimal codeword for it is found which produces the minimum distortion for the fullband case. This codeword completely specifies the appropriate codewords for the corresponding subband filters.

Instead of using the full-band filter in the above algorithm, as a pointer to the subband codewords, lower order filters can be used, derived from the subband signals along the QMF tree. More than one filters would then serve as pointers and each one would define the codewords for a (different) subset of the subband filters. These subband filters would, together, describe the speech envelope in the same frequency region as the corresponding pointer filter. For the purposes of simulation and in order to compare the results of this method to those of the third method, two half-band filters were used as pointers at each time, derived from the signals in the bands 0-2000 Hz and 2000-4000 Hz respectively. The first filter, describing the speech envelope between 0-2 kHz, was used as a pointer for the 4 subband filters corresponding to the frequency regions 0-500 Hz, 500 Hz-1000 Hz, 1000 Hz-1500 Hz and 1500-2000 Hz. (bands 1 to 4). The second filter, describing the speech envelope between 2-4 kHz, was used as a pointer for the 3 subband filters corresponding to the frequency regions 2-2.5 KHz, 2.5-3KHz and 3-3.5KHz (bands 5 to 7). A 7th order filter was used for the lower half-band (i.e. bands 1-4) and 5th order filter was used for the upper half-band (i.e. bands 5-7). The log area ratios (LAR) were used to vector quantize the half band filters. 9 bit codebooks were used for each filter giving a total of 18 bits for the side information. Both transmitter and receiver hold the halfband and subband codebooks as well as the table of associated pairs between each halfband filter codeword and set of corresponding subband filter codewords. The index of the halfband codeword is therefore sufficient to identify the "optimal" subband codewords in each case.

The full search method was used for the m.s.e. vector encoding of the LAR. The use of two half-band filters instead of the full-band filter was dictated from computational considerations. The complexity of an 18 bit codebook is much larger than the complexity of two 9 bit codebooks. The LAR were used instead of any of the Itakura-Saito variants (see chapter 4) in order to provide direct comparison with the scalar coding method of [1]. In [2] a comparison of coding fullband filters using first an LAR measure and then, the Itakura-Saito measure showed no clear subjective preference for any of the two algorithms.

The above structure provides a 9-bit codebook for each of the 7 4th order filters but the total bit rate required for their transmission is only 18 bits per analysis frame instead of 9 x 7 = 63 bits which would have been necessary if conventional 9-bit VQ was used for each subband filter. The penalty paid for the bit reduction is a suboptimal choice of codewords due to the constraints imposed by the selection algorithm. A comparison of the above method of quantization with the third method, in terms of coding efficiency, will be presented later on.

In the third (and last) method to be presented here, the 7 subband filters were encoded independenty from each other. Separate training sets for each subband filter were generated by splitting a large segment of speech into subbands and analysing blocks of speech to obtain 4th order filters for each band. The training set consisted of 111 sec of speech segments (sampled at 8 kHz) from 16 different male and female speakers. No silence blocks were present in the training set of speech segments. This training set was used for all of our codebook training procedures. The frame size for the LPC analysis was 192 samples or 24 msec long. The filter coefficients were transformed to LAR and vector quantized using the full search LBG algorithm with a mean square error criterion. Codebooks of 1-9 bits were generated for the first band (0-500 Hz) and 1-8 bits for all subsequent bands. For the purposes of comparison the LARs were also scalar quantized using the method proposed by Viswanathan and Makhoul [1]. The logarithm of the mean square distortion, over the whole training set, is plotted against the number of bits in figures 6.2-6.8, for both the scalar and vector quantization cases. It can be seen that vector quantization has a constant 5-6 bit advantage over scalar quantization for the bit range from 1 to 15 bits/filter. This is equivalent to an overall bit advantage of 35-42 bits for the filter side information for the VQ method over the scalar quantization method.* The distortion values above were obtained using the same segments of speech that were used in the training phase. The bit advantage indicated above must therefore be considered as an upper limit to the efficiency of VQ coding, although the large amount of training data makes this a fairly representative value. The straight lines in figures 6.2-6.8 show the performance of the second method of coding described above. Since 9 bits were used for 4 bands for the first half of the spectrum, the average number of bits/band is 9/4 ≈ 2 bits/band. From the figures it appears that the resulting distortion from this method of coding is slightly higher in a bits/band basis than conventional VQ of the subbands. Thus it seems that, although for the second method all 4 subbands are encoded together as a vector, which should be able to take advantage of correlations amongst the different filters from each subband, the ensuing operations introduce enough inaccuracies to overcome this advantage. The third method, of directly encoding the subband filters independently was therefore used in all further simulations. This method of encoding the short term envelope of the speech signal offers clear advantages over scalar encoding. In addition, since the allocated bits are split amongst 7 codebooks of small dimensions, the complexity is very low compared to the equivalent VQ encoding of the fullband signal.

Since the subband filters are encoded separately, some criterion has to be applied in order to determine the bit allocation amongst the different subband filters. The criterion chosen was that the spectral distortion should be approximately constant over the whole frequency region. This criterion is usually

*The above advantage is consistent with the assumption that significant (linear and non-linear) correlations exist among the subband filter parameters.

observed when full band LPC analysis and coding are performed. This is known to perform well and conform to subjective criteria for the telephone band speech. The aim was, therefore, to find that bit distribution which minimized the sum of the mean square LAR distortion from each band, uniformly weighted, subject to a particular total number of bits. The distortions/band are replotted on a linear scale in figures 6.9-6.15. (crosses). A comparison of the 7 curves reveals that the distortions in each band for any fixed number of bits are unequal and that they decrease as the center frequency of the bands increases. This can be attributed to the fact that as the frequency increases the spectral activity (i.e. the number of different possible shapes) decreases. Sixth order polynomials were fitted to these points as shown by the continuous lines in figures 6.9-6.15 using a least squares routine (E02ACF) from the NAG library [3]. These polynomials define functions of distortion versus bit rate for each band, of the form:

$$d(k,b_k) = f_k(b_k) \qquad 6.8$$

where d is the distortion, k the band index (1-7), b_k the number of bits allocated for band k and $f_k(b_k)$ a sixth order polynomial in b_k.

The aim is to minimize

$$D = \sum_{k=1}^{7} d(k,b_k) \qquad 6.9$$

subject to

$$B = \sum_{k=1}^{7} b_k = \text{constant} \qquad 6.10$$

and

$$1 \leqslant b_k \leqslant 9 \text{ for band } 1 \qquad 6.11a$$

$$1 \leqslant b_k \leqslant 8 \text{ for bands } 2\text{-}7 \qquad 6.11b$$

The upper limits for the codebook sizes were chosen on the basis of complexity considerations. To find the minimum solution $\{b_k\}$ another minimization routine from the NAG library was used [3] [E04UAF]. Since the functions $f_k(b_k)$ were given as continuous and the routine was geared for non-integer programming the resulting bits/band b_k were real numbers. These were simply rounded up or down to the nearest integer value. If the average bit rate constraint of 6.10 was violated, bits were added or subtracted from the band allocations such that this resulted in the minimum deviation from the optimum real number allocation. The bit distributions for 18, 30 and 40 total number of bits is given in table 6.T1.

bits/band	1	2	3	4	5	6	7
18	5	4	3	3	1	1	1
30	6	6	5	4	3	3	3
40	7	7	6	6	5	5	4

Table 6.T1 Optimal bit distributions for the subband filters

Note that for a total number of bits equal to 18, the bit rate for the filters was the same as for the second coding method. These values were used to encode the filters using methods two and three. With all other parameters the same there was about 1 dB advantage in SNRSEG when method three was used, over method two.

The bit rates/band given in table T6.1 provide an optimum fixed bit allocation. Variable bit allocation could possibly lead to improved results. To the knowledge of the author no formula has been proposed to relate the speech power/band to the m.s.e. resulting from the coding of the LPC parameters. Variable bit allocation based upon the individual spectral distortions (LAR) in the same way that was used to obtain the fixed bit allocation would result in increased complexity and require additional side information.

In a series of experiments that were carried out, the assumption was made that the noise power due to LPC spectral errors was related to the speech power through a relationship of the form

$$\sigma_{qk}^{2} = A\ (\sigma_{xk}^{2})^{n}\ 2^{-2R_k} \qquad 6.12$$

A reasonable range of values for n was investigated but no value was found to produce a quality better than the fixed allocation of 6.T1. The coded speech was, in general, more burbly.

An improvement over the fixed bit allocation of 6.T1 could be achieved as follows: Various patterns of fixed bit allocations could be used and the coded speech obtained could be compared subjectively to determine the optimum coefficients. With the assumption of the bit allocation of 6.T1 used as a starting point and that each band allocation is allowed to deviate from the starting point by ± 1 bit, the number of possible combinations can be calculated as follows: Assume that one of the bands is given an extra bit. To maintain the same total number of bits, the number of bits in one of the remaining six bands must be reduced by one. There are therefore six possibilities. Since the first band chosen can be anyone out of the 7, the total number of combinations is 7 x 6 = 42! for a one bit deviation from the starting configuration. Clearly, optimizing the bit allocation in this fashion is quite unpractical. Therefore, the fixed bit allocation dictated by the average distortion per band curves and the optimization routine with an overall minimum distortion over the spectrum, was used for all further simulations.

6.3 SIDE INFORMATION: GAINS

Three more parameters are additionally needed to complete the set required as side information for this coder structure: First, since the pulses in a multipulse algorithm are encoded using block companded quantization (PCM-AQF), the variances of the pulses need to be transmitted. The pulse amplitudes are normalized by this estimate and subsequently coded by a unit variance memoryless quantizer matched to the pdf of the pulses in each band. Second, the use of variable bit allocation requires some prior knowledge about the noise variance in each band. In a predictive structure this is usually assumed to be proportional to the LPC residual variance, hence the power of the residual needs also to be transmitted. Finally, since variable bit allocation at low bit rates would lead to zero bit quantization in order to preserve the short signal spectrum, noise of the right variance needs to be injected into the empty bands. Therefore the variance of the LPC residual is also needed for reproducing the short term spectrum of the signal. This follows from the decision to transmit the LPC envelope even for the "empty" bands in order to preserve the signals envelope with as much fidelity as possible.

Therefore the three parameters can be narrowed down to two, the variance of the excitation pulses and the LPC residual of the subband signals.

The variance of the noise power injected into the empty bands is usually a fraction of the LPC residual power. This is because the empty bands represent the valleys of the signal in the frequency domain, where amplitude discrimination by the auditory system is poor. By reducing the injected noise power, inaudible distortion is introduced in the speech signal whilst the "noisiness" perceived due to the aperiodicity of the signal injected is reduced. This can be considered equivalent to postfiltering operations for noise reduction in full band coding of speech [4,5].

The variance of the subband signals for the requirements of the bit allocation algorithm can also be substituted by some other parameter which is itself correlated to the variance of the subband signals. Then the generalized form of the bit allocation formula derived in appendix D, can be used, once the relationship between the error power per band and these parameters can be determined experimentally. This can be achieved by measuring the parameter's variance to noise ratio in each band individually (assuming parameters with zero means). Such measurements were made and applied quite successfully and are presented in a following section.

The common use bit allocation formula of 3.2-162 is based on the assumption that the quantizer performance is described by a *formula* of the form

$$\sigma_q^2 = \epsilon_*^2 2^{-2R} \sigma_x^2 \qquad 6.13$$

where σ_q^2 is the subband noise variance and σ_x^2 is the LPC residual, when predictive coding is used. Since this is not necessarily so for the analysis by synthesis coding used here, the choice of parameters $\{\sigma_x^2\}$ need not be the variances of the residual/band. It was therefore decided to use the variances of the pulses as the set of parameters for the side information. This provides a good AQF quantizer for the pulse amplitudes, minimizing any variance mismatch for the block adaptive (unit variance) quantizer. The problem that immediately arizes from such a choice is that the variances of the pulses cannot be known before the pulses are found. Since the number of pulses to be used in each band will depend on the results of the bit allocation formula and the bit allocation formula requires these variances in order to determine how many bits, hence how many pulses to be allocated, a boot-strap situation arizes.

The way this problem was overcome was to use some other parameter in the bit allocation formula, in this case the LPC residual powers, and then use these parameters, in a training phase, to find how many pulses, on the average, are allocated in each band for a large amount of training speech data, and a particular overall bit rate. Following this technique, after the average number of pulses/band were found, the coder was run using the variances of the average number of pulses as its side information.

The actual number of pulses finally used for the excitation in each band will not, in general, be the same as the average number of pulses obtained as above. Therefore a variance mismatch will arize. To assess the extent of this variance mismatch the following procedure was used:

Assume that for a particular band the average number of pulses $(P_m(i))$ was found to be L where $L \leq M$, the multipulse frame. L is a function of the subband index, different for different bands. The rms of the number of pulses (model r.m.s.) is given by

$$RMS_{m1} = \sqrt{\frac{1}{L} \sum_{i=1}^{L} P_m^{\ 2}(i)} \qquad 6.14a$$

or

$$RMS_{m1} = \sqrt{\frac{1}{L} \sum_{i=1}^{M} P_m^{\ 2}(i)} \qquad 6.14b$$

where, in the second form (6.14b), zero amplitude pulses are assumed to occupy locations in the frame for which no pulses are allocated.

Another useful parameter closely related to the above is given by

$$RMS_{m2} = \sqrt{\frac{1}{M} \sum_{i=1}^{M} P_m^{\,2}(i)} \qquad 6.15$$

RMS_{m2} can be thought of as an approximation to the rms of the LPC residual in as much as the pulses {P(i)} are an approximation to the LPC residual itself, for the purpose of synthesising the encoded speech (subband) signal. Note that, both the zero amplitude and the non-zero amplitude pulses together, model the excitation subband signal.

In a coding situation, one such value of RMS_{m1} or RMS_{m2} is obtained for each band, at the encoder. The 7-dimensional vector comprized of these values is coded (the coding of the vector is described in a later section) and used at both the encoder and decoder to derive an "optimum" pulse allocation. The pulse allocation algorithm assigns a different number T of non zero pulses for each band.

The rms of the allocated pulses $\{P_a(i)\}$ is given by

$$RMS_a = \sqrt{\frac{1}{T} \sum_{i=1}^{T} P_a^{\,2}(i)} \qquad k \leqslant M \qquad 6.16a$$

or

$$RMS_a = \sqrt{\frac{1}{T} \sum_{i=1}^{M} P_a^{\,2}(i)} \qquad 6.16b$$

where T is a function of the subband index (i.e. it is different for different bands).

The relation between L and T is that the expected value of T,E(T), is (approximately) equal to L:

$$E(T) = \frac{1}{N} \sum_{n=1}^{N} T_n = L \qquad 6.17$$

where N is the number of frames in the training data.

To assess the degree of variance mismatch an estimate of the average difference between RMS_{m1} and RMS_{m2} (the model rms), and RMS_a (the rms of the pulses from the final pulse allocation) is needed. Since these values are expected to have a large variance themselves, a relative error measure is more appropriate. The measure used was:

$$\text{ERROR} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \frac{(RMS_a - RMS_m)^2}{(RMS_a)^2}} \qquad 6.18$$

where RMS_m is either RMS_{m1} or RMS_{m2}. N is again the total number of frames in the training data. Both RMS_a and RMS_m are functions of the time frame and the subband index. Typical values for one speech file obtained for a model pulse distribution appropriate for a bit rate around 4.8 kb/s are given in table 6.T2. Note that subband frames for which T in 6.16 was zero were not taken into consideration for the calculation of L or ERROR and the value of N was modified accordingly for these bands.

BAND	1	2	3	4	5	6	7
L	10	3	2	2	2	2	2
RMS_{m1}	0.28	0.21	0.16	0.15	0.15	0.15	0.21
RMS_{m2}	0.38	0.58	0.65	0.65	0.65	0.65	0.64

6.T2 Relative error between two model rms values and the true rms value

Although the resulting errors are relatively small in both cases, RMS_{m1} is clearly preferable to RMS_{m2}, as RMS_{m1} is a better approximation to RMS_{a}. In practice though the SNRSEG performance from using the two parameters was the same within a few tenths of a dB and, perceptually no clear preference existed between the two, when used in the AQF quantizer.

6.4 BIT ALLOCATION

The bit (or pulse) allocation formula involves a relationship between the resultant noise power per band K, σ_{qk}^2, which is not known a priori and a side information parameter for each band k termed σ_{xk}^2. The general assumption made here is that a linear relationship exists between the logarithm of the ratio $\sigma_{qk}^2/\sigma_{xk}^2$ and the number of bits allocated in that band, R_k i.e.

$$\log \frac{\sigma_{qk}^2}{\sigma_{xk}^2} \;\alpha\; a_k R_k + b_k \qquad 6.19$$

or by redefining the constants:

$$10\log_{10} \frac{\sigma_{xk}^2}{\sigma_{qk}^2} = (-2a_k R_k)\log_{10}2 + 10\log\epsilon_{*k}^2 \qquad 6.20$$

which can be written in a more familiar form

$$\sigma_{qk}^2 = \epsilon_{*k}^2 \, 2^{-2a_k R_k} \sigma_{xk}^2 \qquad 6.21$$

How the actual number of bits is translated into a number of pulses will be described shortly. In order to assess the applicabiity of 6.21 for the multipulse algorithm, the parameters RMS_{m1} and RMS_{m2} were used for σ_{xk}^2 and the left-hand-side of 6.20 was plotted against the number of bits R_k for different bands for one speech file (male). The results for RMS_{m1} are shown in figure 6.16 and those using RMS_{m2} are shown in figure 6.17. Figure 6.18 shows the results obtained when σ_{xk}^2 in 6.20 represents the speech power/band. The left-hand-side of 6.20 in each case represents a time average as in SNRSEG. The straight line corresponding to a slope of 6 dB/bit is also shown for comparison. Each band was coded independently using the multipulse algorithm of [6]. This algorithm was used since it possesses low complexity. It is known that higher complexity algorithms fail to show any significant perceptual improvements at low bit rates.

Formula 6.21 contains two band-specific constants, ϵ_{*k}^2 and a_k. The common use bit allocation formula of 3.2-162 is derived under the assumption that $\epsilon_{*k}^2 = \epsilon_*^2$ = constant and that $a_k = 1$. The condition $a_k = 1$ is translated into the 6 dB per bit rule. It can be seen from figure 6.16 that in using RMS_{m1} for σ_{xk}^2 in 6.20 (or 6.21) this situation is approximately true. This is not so for figure 6.17 where the curve for the first band is very different from the curves for the rest of the bands and therefore ϵ_{*k}^2 = constant cannot be assumed.

The more general solution to the bit allocation problem based on 6.21 under arbitrary and unequal ϵ_{*k}^2 and a_k was derived and is presented in appendix D. The appropriate constants ϵ_{*k}^2 and a_k for each band were calculated from the data of figures 6.16 and 6.17 using straight line fitting routines. Although the new bit allocation formula provided a marked improvement in the performance of the coder using RMS_{m2}, no perceptual advantage was obtained for RMS_{m1} confirming the results of figure 6.16.

In 6.21 R_k is the equivalent bit rate/sample. In multipulse coding (a delayed decision algorithm) though each sample is not coded individually but, rather, bits are allocated on a block (frame) basis. Assuming M samples/block, R_k can be replaced by $P_k = MR_k$, the number of bits/block. Formula 6.21 can now be modified to

$$\sigma_{qk}^{2} = \epsilon_{*k}^{2}\, 2^{-\frac{P_k}{M}}\, \sigma_{xk}^{2} \qquad 6.22$$

For simplicity ϵ_{*k}^2 = constant and $a_k = 1$, although the basic assumptions also hold for the general case as shown in appendix D.

Ramstad's algorithm (see Appendix C) can be modified to operate on a block basis as follows:

In the conventional formula, the assumption that the difference in bits allocated to two bands a and b is 1 (see equation C4, appendix C). i.e.

$$R_a - R_b = 1 \qquad 6.23$$

leads to the result

$$\frac{1}{2}\log_2 \frac{\sigma_{xa}^{2}}{\sigma_{xb}^{2}} = 1 \qquad 6.24a$$

or

$$\sigma_{xa}/\sigma_{xb} = 2 \qquad 6.24b$$

which implies that halving σ_{xk} is equivalent to adding one bit to band k. The algorithm was designed to deal with the constraint of integer bit allocation i.e.

$$[R_a - R_b] = \text{integer} \qquad 6.25a$$

and

$$\min [R_a - R_b] = 1 \qquad 6.25b$$

Since the bits in the present scheme are allocated on a per block basis, only the total number of bits/block need be an integer, hence one can afford to use:

$$\min [R_a-R_b] = \frac{1}{M} \qquad 6.26a$$

where M is the multipulse frame size in samples and

$$R_a - R_b = m \frac{1}{M} \qquad 6.26b$$

where m is an integer.

this leads to

$$\min [P_a-P_b] = M(R_a-R_b) = M \frac{1}{M} = 1 \qquad 6.27a$$

and

$$P_a-P_b = m \qquad 6.27b$$

i.e. an integer bit allocation/block. Since

$$R_a-R_b = \frac{1}{2} \log_2 \frac{\sigma_{xa}^2}{\sigma_{xb}^2} \qquad 6.28$$

for

$$R_a - R_b = \frac{1}{M} \qquad 6.29$$

then

$$\frac{1}{M} = \frac{1}{2} \log_2 \frac{\sigma_{xa}^2}{\sigma_{xb}^2} \qquad 6.30a$$

and

$$\sigma_{xa}/\sigma_{xb} = 2^{\frac{1}{M}} \qquad 6.30b$$

Therefore a reduction in the bit rate per sample R_k by 1/M bits (or, equivalently, a reduction of the bit rate/block P_k by 1 bit) is equivalent to reducing σ_{xk} by a factor $2^{1/M}$.

The modified Ramstad algorithm can therefore be described in terms of the following steps:

1. Start from the maximum σ_{xk}. Allocate 1 bit/block to this band and divide σ_{xk} by $2^{1/M}$.
2. Subtract 1/M bits from the total bit rate calculated on a per sample basis i.e. from $R = \Sigma R_k$ or, equivalently, subtract 1 bit from the total bit rate calculated on a per block basis i.e. from $P = \Sigma P_k$.
3. If R (and P) now equal zero exit else goto 1.

Noise weighting can be applied as in appendix C by replacing each σ_{xk}^2 by $(\sigma_{xk}^2)^w$. It was found that the best results, perceptually, were obtained when no weighting was used.

6.5 PULSE ALLOCATION

Once the bits have been allocated they have to be translated into a number of pulses. For the coding of the pulse amplitudes 3 or 4 bits can be used according to the overall bit rate. For operation around 9.6 kbs/sec 4 bits per pulse are appropriate whereas for 4.8 kbs/sec 3 bits/pulse are used. In addition, for each block, the pulse positions have to be transmitted. The required number of bits needed is given by the following equation

$$B = \log_2 \frac{n!}{e!\ (n-e)!} \qquad 6.31$$

where n is the (multipulse) frame size and e is the number of allocated pulses. B is rounded up to the nearest integer. In effect, the relationship between bits allocated and number of pulses is given by:

$$B = 3*e + \log_2 [\frac{n!}{e!(n-e)!}] \qquad 6.32$$

for 3 bit coding, and

$$B = 4*e + \log_2 [\frac{n!}{e!(n-e)!}] \qquad 6.33$$

for 4 bit coding, where [ ] denotes rounding up to the nearest integer.

Setting the framesize to 16 msec, results in 16 samples/frame for each subband, whereas for the fullband case this results in 128 samples/frame. A plot of B versus e in equations 6.32 and 6.33 is shown in figures 6.19 and 6.20 for n = 16 samples and n = 128 samples respectively. These relationships are also given in tables 6.T3-6.T6 where the bit rate in bits/sec and also the number of bits per pulse is given. Various conclusions can be deduced from the graphs: The fullband graph of figure 6.20 shows an almost linear relationship between the number of bits and the number of pulses, which in turn implies an *almost constant ratio* of bits/pulse as can be seen from tables 6.T5 and 6.T6. (for 1 up to 16 pulses). Also, the difference between 3 bits/pulse and 4 bits/pulse is small for moderate bit rates (~ 50 bits). By contrast, in the subband case, figure 6.19 shows a highly nonlinear relationship, where the number of bits/pulse decreases significantly with the number of pulses (see tables 6.T3 and 6.T4). Also the difference between 3 pulse and 4 pulse quantization for the amplitudes is significant compared to the fullband case and for the same number of bits.

A comparison between the two graphs reveals that for the same number of bits overall, more bits are allocated for position coding rather than the amplitude coding in the fullband case as opposed to the subband case. It follows that quantization of the amplitudes will be more critical in the subband case and a larger drop in SNR would result when the amplitudes are coded in the subband scheme as compared with the fullband scheme. This was found to be true in practice.

For the purposes of bit allocation the graph in figure 6.19 is misleading since it shows a smooth relationship between number of pulses and number of bits, as opposed to the true, staircase relationship shown in figure 6.21. It can be seen that the transition levels are quite widely spaced, especially at low bit rates. This is translated into an unavoidable inaccuracy in the pulse allocation algorithm. The situation that arises is that not all the bits allocated to a particular band can be translated into pulses, especially at low pulse (bit) rates. Therefore the problem of allocating an integer number of pulses arises, very analogous to the integer bit number constraint of conventional bit allocation algorithms.

The overall pulse allocation algorithm that was employed is shown in figure 6.22 in a flowchart form. The gains vector (which can be modified for the purposes of noise shaping) is used to determine the bit vector (whose elements are the bit allocations per band), through a modified Ramstad algorithm as was described earlier. Next, the number of bits for each band is translated into a provisional number of pulses according to the inverse of equation 6.32 or figure 6.19. In addition, the number of bits left over from each band, due to the integer pulse allocation constraint, is stored in another vector (called the "excess" vector in the flow chart) whereas the additional number of bits required to increase the provisional number of pulses by one in each band is stored into the "required" vector. These are fed into a reallocation routine which allocates the extra bits in the "excess" vector to readjust the number of pulses/band and allocate as many of the available bits as possible. A summary of the reallocation routine is as follows: Let the total number of bands with less than 16 provisional pulses (=multipulse frame) allocated be equal to M. Let the component i of the "excess" vector be ΔE_i and the corresponding component i of the "required" vector be ΔR_i. The algorithm minimizes the sum

$$SUM = \sum_{M} (|\Delta E_i| + |\Delta R_i|) \qquad 6.34$$

through an iterative procedure. The details of the algorithm can be found in appendix G.

6.6 GAINS CODING

The side information parameters (gains) presented by 6.14 or 6.15 have to be coded (quantized) prior to transmission. There is one such parameter per band k denoted here by G_K. For each time block of data therefore a vector parameter $\{G_K, k=1,2...,7\}$ exists which to a certain extent, is an estimate (biased or unbiased) to the vector formed by the standard deviation of the LPC residual per band. (The term "gains" is borrowed from LPC vocoder literature).

An efficient way to code the vector G_K is through vector APCM or block vector encoding in an AQF mode: The vector components are normalized by the value given by:

$$P = \sqrt{\sum_{k=1}^{7} G_k^{2}} \qquad 6.35$$

to obtain a new vector given by $\{G'_K, k=1,....,7\}$ or $\{G_K/P, k=1,....,7\}$. The length of the new vector is always equal to 1 since

$$\text{Length} = \sqrt{\sum_{k=1}^{7} G'_k{}^{2}} = \frac{1}{P}\sqrt{\sum_{k=1}^{7} G_k^{2}} = 1 \qquad 6.36$$

Therefore if the vector $\{G_K\}$ can be considered to represent a point in a k-dimensional space, the normalization by P confines each new point represented by $\{G'_K\}$ to the surface of the unit hypershere in the k-dimensional space. Since the possible locations of the vector to be encoded have been drastically reduced from the whole of the k-dimensional space to the surface of the unit hypershere, much fewer bits are required to encode the vector $\{G'_K\}$ than to encode the original vector $\{G_K\}$. The normalization can also be considered as splitting the vector $\{G_K\}$ into a parameter P representing its length and $\{G'_K\}$ representing its orientation in the k-dimensional space.

Parameters P and $\{G'_K\}$ are expected to have linear and non-linear dependencies between them and the most economical way, in terms of bits, to encode them would perhaps be by encoding the augmented vector $\{G'_K,P\}$ The problem that arises is that of a choice of an appropriate distortion measure: The value of P would generally demand more encoding accuracy than any of the individual components of $\{G'_K\}$ which implies some form of weighted distortion measure. A product (gain-shape) codebook structure could also be used. For simplicity it was decided to encode P and $\{G'_K\}$ separately, using a scalar codebook for P and a vector quantizer for $\{G'_K\}$.

Parameter P can be coded in a variety of ways. It was decided to use a logarithmic compression and uniform quantization of the logarithmically compressed value. The base of the logarithm used is of course unimportant. Logarithmic quantization of the parameter P (the overall gain) which is an estimate of the fullband signal's standard deviation is a reasonable choice since the intensity jnd is proportional to the intensity, as was mentioned in chapter 2. The log-linear type compressor used in log-PCM was avoided because 16 bit representations of the original speech were available, of which, the 4 least significant did not contain much (audible) information about the signal. A log-linear type compressor would have placed an inappropriate degree of accuracy on low signal values of a range around the 4 least significant bit representations.

Five bits were used for uniform encoding of the logarithmically compressed P. This value could be lowered to 4 bits if the range of the input signal was confined to 12 bits (which is considered as adequate for most applications). The value of 5 bits leads to an almost imperceptible degradation of the signal compared to the uncoded case. The update of P (and $\{G'_K\}$) was done every 16 msec. Further bit savings can be made by differentially encoding P: Successive values are highly correlated as would be expected. A first order long term autocorrelation coefficient of about 0.99 was obtained for the logarithmically compressed value of P.

The vector $\{G'_k\}$ was encoded using the full search LBG algorithm. In a first attempt, an m.s.e. criterion was used directly on the components for $\{G'_k\}$. This was found unacceptable since it distributes the distortion equally amongst the vector components. This implies that when one or two components are dominant in the sense that their values are much higher than the rest, the distortion in the smaller components can actually exceed by many times the component values themselves. Consider figure 6.23a,b,c,d. The thick line represents 4 spectral envelopes from successive speech blocks of the uncoded signal. The envelope was obtained through a 12th order autocorrelation analysis. The thin line represents 4 envelopes from the coding noise obtained by subtracting from the 4 time domain blocks of uncoded speech signal the corresponding time-domain blocks of coded signal. Again a 12th order autocorrelation analysis was performed on the noise signal. Although the noise signal cannot be considered as an all-pole signal the envelopes obtained in this way were accurate enough for our purposes. The coded signal used to obtain figure 6.23 was processed through the algorithm described in this chapter, but with the gains side information left uncoded. It can be seen that the envelope of the noise remains under the speech envelope throughout. Figure 6.24(a,b,c,d) was obtained as above but with the gains encoded using a m.s.e. on the components of $\{G_k\}$. It is clear that the level of noise in the higher frequency range is unacceptable. Nine bits were used for the codebook.

Figure 6.25a,b,c,d represents the same blocks with each component of the gains vector logarithmically compressed prior to vector encoding with a m.s.e. criterion. It can be seen that the resulting envelope distortion is much smaller than before. Nine bits were used for constructing the new codebook. A comparison between a file with coded and uncoded vectors revealed a small but perceptible degradation. Vector prediction is expected to be beneficial in encoding $\{G'_K\}$ as well. The correlation (in the spectral envelopes) of successive speech frames is clearly demonstrated in the above figures.

6.7 PULSE AMPLITUDE CODING

The pulse amplitudes were coded in an AQF mode: The amplitudes in each band were normalized by RMS_{m1} or RMS_{m2} and then encoded using either a unity variance Gaussian quantizer or an optimum quantizer. At the decoder, the quantized values were scaled back using the inverse of the normalization factor that was used at the transmitter side.

The optimum quantizer was obtained through the Max-Lloyd procedure implemented using the LBG algorithm to create one dimensional vector codebooks. A different codebook was trained for each band. The training data used were the unquantized (normalized by RMS_{m1} pulse amplitudes, obtained during the analysis phase of the algorithm. The pdfs of the pulse amplitudes for each band are shown in figures 6.26a-g. Less data was available for the training phase for the higher bands as opposed to the lower bands because the pulse allocation algorithm allocates more pulses to the lower than to the higher bands. Blocks with zero pulse allocation were of course excluded from the calculation of the pdfs. The total number of training data used for each band is given below

BAND	1	2	3	4	5	6	7
Number of pulses	44315	10954	3684	2188	1552	943	772

Table 6.T3: Total number of training data per band

The number of bins was obtained by dividing the total range of pulse amplitude values by the square root of the corresponding number in table 6.T3. The superimposed curves describe the Gaussian distribution given by

$$N(x,\sigma_x^2) = \frac{1}{\sqrt{2\Pi\sigma_x^2}} \exp\left[\frac{-(x-\overline{x})^2}{2\sigma_x^2}\right] \qquad 6.37$$

The solid curve describes the distribution N (0,1) whereas, for the dashed curve the distribution is $N(x,\sigma_x^2)$ where x and σ_x^2 represent the mean and variance of the actual training data. A comparison between the solid and dashed curves shows the degree of variance mismatch when a unit variance gaussian quantizer is used for coding. It can be seen that the mismatch is fairly small.

The distributions can be seen to diverge from a Gaussian towards a bimodal distribution as the subband index is increased. This is also reflected in the quantizers designed with the training data of fig. 6.26a-g using the LBG algorithm. Fig. 6.27b-h shows the optimum quantizers for bands 1-7. Fig. 6.27a shows a gaussian quantizer for comparison. From figure 6.27 it can be seen that the mid-range step sizes grow smaller as the subband index is increased at the expense of the low range step sizes. The SNRSEG results from using the two types of quantizer (Gaussian and optimum) were the same usually to within a tenth of a dB. This was perhaps to be expected, since the optimum quantizer departs from a Gaussian type when the allocated number of pulses is small, which in turn implies a low signal power in the relevant band. It follows that the resulting quantization noise variance from this band will also be small since it is somewhat proportional to the speech power in that band. Also, informal listening tests revealed no preference to coding with a Gaussian or an optimum quantizer. This perhaps reflects the fact that the locations of the pulses are more crucial than their actual amplitudes for low pulse densities.

The multipulse algorithm used to encode each band separately, (once the allocated number of pulses were known), was that of reference [6], employing a sequential (iterative) search with no intermediate, overall amplitude optimization. The autocorrelation approximation was used to substitute for all autocovariance estimates. Full details of the algorithm are described in appendix F.

6.8 PREFERRED SIMULATION PARAMETERS FOR A 4.8KBITS/SEC CODER

A three stage QMF tree was used for the subband splitting. The same 32-tap filters were used at each stage. The filter coefficients were obtained from table 3.2-T7 (chapter 3). The eighth band was not encoded at all, and zeros were used in the synthesis QMF in place of the eighth band. The resulting sampling frequency for each subband signal was 1KHz. The LPC frame size used was 32 samples (32msec). The -LPC coefficients were obtained using the Burg method. This analysis technique is particularly attractive here, considering the small number of samples used in the derivation of the coefficients. Fourth order filters were used in each band, vector quantized, with the number of bits given by table 6.T1 for a total number of bits equal to 30. $\{RMS_{m,i}\}$ of 6.14 was used to represent the gains obtained every 16 samples (16 msec). Two such frames (which coincide with the multipulse analysis frame, again of 16 msec duration), comprised an LPC frame. The update rate for both the LPC frame and the multipulse (gain) frame was the same as the corresponding frame duration. No block overlapping was used. To obtain $\{RMS_{m,i}\}$ the average number of pulses allocated per band were estimated in a training phase. These values are shown in table 6.T2 (L). To obtain L the LPC residual was used as the side information to the bit allocation algorithm. Having estimated L for each band, RMS_{m1} for each band was obtained using equation 6.14, after the L pulses were found by running the multipulse algorithm through a dummy run for each block.

The multipulse algorithm, used both in the dummy run to determine the gains and in the actual coding was that described in the previous section. The gains were normalized by the overall rms value which was coded using 5 bits whereas the normalized gains vector was coded with 9 bits. (With log compression in both cases).

Three bit Gaussian quantizers were used for coding the normalized by RMS_{m1} pulse amplitudes. For the "empty" bands, where no pulses have been allocated, white noise was used at the receiver side of the algorithm to drive the LPC filters. The rms of the noise was the same as the value of RMS_{m1} for the corresponding band, multiplied by a "post-filtering" factor of around (0.5-0.7). Forty seven bits were used for coding the pulse amplitudes and positions, bringing the total bit rate to 4.8 kb/s. These were distributed amongst the bands according to the pulse allocation algorithm, for which the front end section was the modified Ramstad method mentioned above. A Ramstad ratio of 2 was found to give the best performance perceptually (i.e. equivalent to no noise shaping). Increasing values gave a more "burbly" quality to the coded speech whereas for smaller values than 2 a "whispering" image effect could be heard in the background.

Training data for the codebooks were derived using speech segments from 16 different speakers. The number of training vectors used for the LPC codebooks was 3487, giving a minimum training vector number versus codebook level ratio of 3487/64≈55 which must be considered as adequate. The number of training vectors for the gains codebook was 6975 giving a training vector number versus codebook level ratio of 6975/512≈14. This number is considered to be rather low and better codebooks can be expected for a larger number of training data.

The noise shaping factor g (equation F1 of appendix F) was kept at 0.8 for all bands. No attempt was made to evaluate the effect of changing g.

6.9 SPEECH QUALITY

To assess the rate compression effectiveness of the algorithm, encoded speech was compared with fullband encoded speech using the multipulse algorithm of [6] at 4.8 kbits/sec. The amount of roughness in the signal coded with the proposed algorithm was much less than the one coded with the fullband algorithm especially for female speech. This was also reflected in the SNRSEG results which were, in general, 2-3 dB higher for the proposed algorithm. The overall speech quality in the subband algorithm was degraded though because of click and "knocking" sounds which although of a very low level were perceptually annoying. Slight burbling was also evident in certain sounds. These effects were thought to be generated from blocks of speech where an insufficient number of pulses were allocated, since, even with unquantized parameters most effects were still evident. Blocks of speech in which only one or two pulses were allocated gave rise to gaps in the synthesised subband signals. To investigate the distortion effects further, an adaptive algorithm was designed to locate and fill the gaps in the subband signals with appropriately scaled noise samples. A significant reduction in the burbling type of distortion was obtained but the other types of distortion remained present. A similar result was obtained when the coded through the proposed algorithm speech signal was then further encoded through the full-band multipulse algorithm which automatically generates scaled noise in the higher frequency spectrum.

Since the click and "knocking" distortions persisted it was decided that these effects were due to the interruption of the periodicity of the speech signal in the subbands and not just due to gaps in the subband signal. This would explain why no improvement was obtained over certain types of distortion when the gaps were filled with white noise.

In addition to the above effects the choice of a model pulse rms such as RMS_{m1} or RMS_{m2} for the gains proved difficult to handle since, when any of the parameters were changed in value, there was a resulting change in the pulses used in 6.14 to obtain the model RMS. It was found necessary to obtain a new codebook each time any of the parameters were changed, since the old codebook was no longer representative of the new $\{RMS_{m1}\}$ vectors.

6.10 A NEW ALGORITHM

An important drawback of subband algorithms in general is that the subband signals are encoded independently, although significant linear and nonlinear correlations exist between bands. These can be split into four categories related to the types of parameters used to model the signal in the above algorithm. The first correlation relates to overall short term power variations. This is taken care of by normalizing the gains vector by the overall rms. In addition, the normalized subband powers are interdependent. Vector quantization of the gains vector exploits these redundancies with significant savings in terms of bit rate.

Dependencies also exist amongst the LPC filters from different bands. A straightforward way to take advantage of these particular redundancies is to use fewer vectors of larger dimensions whose components derive from more than one LPC filter. Since the bit allocation for the LPC filters is based on the minimization of the same distortion criterion used in the VQ of the LAR, the same LPC bit allocation algorithm can be used on the combined vectors. The penalty from this approach is a significant increase in complexity.

The last redundancy is related to the pitch structure within each subband: The multipulse algorithm itself does not involve any pitch prediction. This problem is accentuated by the fact that the same pitch information needs to be specified in all the subbands. Encoding all the subband signals individually wastes bits, in order to provide what is essentially the same pitch information, to each band individually.

One way to conserve bits is to place constraints upon the possible locations for the pulses in the subband excitation signals. An analytically tractable way to do this is to constrain the pulses in all the bands for which an excitation signal is to be transmitted (the "passbands"), to occupy the same locations with respect to the start of the frame. Therefore, only one set of locations needs to be specified for all bands. In order to avoid the need for extra side information, the number of pulses per band must also remain constant across the bands. Since bits for the pulse amplitudes are still needed for each band there is a limit to the maximum number of passbands, whilst the actual number per frame can be made adaptive.

Let the error power between the coded and original subband signal in band j and for I pulses be P_{jI}. Instead of minimizing P_{jI} separately for each band the overall error power is formed:

$$P_I = \sum_{j=1}^{M} P_{jI} \qquad 6.38$$

where M is the total number of passbands for the current frame and (j = 1,2,....M), the passband indices.

Instead of selecting each pulse location to minimize P_{jI} for band j, the pulse location m_I can be chosen which minimizes P_I in 6.38. The pulse amplitudes can be optimized individually for each band, as will be seen in section 6.10.2.

An iterative algorithm can be employed as before where each additional set of pulses (related to only one location m_i and several bands) can be chosen to sequentially minimize P_I.

At each stage of the minimization process there will be dominant bands j for which P_{jI} will be large. These bands will mostly determine the chosen pulse location. As more pulses are introduced, the error P_{jI} for the previously dominant bands will be reduced and other bands will now become dominant, therefore, pulse locations which are good choices for all the bands will eventually be found, given enough bits. Note that since the amplitudes are individually optimized for each subband a monotonic decrease in distortion in all bands is guaranteed.

Assume that for the first few pulses one dominant band exists: A schematic graph of distortion versus pulse rate is shown in figure 6.28a for this band. For those bands which have a high correlation with the first dominant band, the first few pulse locations will also be good choices, although the distortion will be expected to reduce at a slower rate (figure 6.28b). For those bands with low correlation with the dominant band, the first few pulses may reduce the distortion by very little (fig. 6.28c) and only when the noise power in the previously dominant band is reduced sufficiently, the bands with low correlations with the first, previously dominant band, will receive pulses whose locations are good choices in terms of minimizing the error power in these bands.

6.10.1 Choice of Passbands

Due to the limited number of available bits not all bands can be included in the pulse allocation minimizing P_I in 6.38. The spectrum is therefore divided into a number of passbands and stopbands. A simple but effective way to implement a decision algorithm is to use the same bit allocation routine as in the previous coder structure (section 6.4). This algorithm (modified Ramstad) divides the available bits amongst the bands in order to minimize the total mean square error. It also divides the spectrum into passbands and stopbands via a threshold rule: bands that receive (a positive number of) bits are considered as passbands whereas the remaining bands are considered as stopbands.

The same procedure can be followed here although an important modification must be implemented: Since all the passbands are modelled with the same number of pulses there must be a minimum number of pulses that can be used, related to the minimum pitch period that can be encountered. A reasonable number is 3 pulses for a frame size of about 16 msec. For a particular band to be considered as a passband it must receive, in the bit allocation formula, at least (3 pulses) x (3 bits each for the amplitudes) = 9 bits plus a small number of bits for its contribution to the pulse position coding.

Therefore a threshold value for the allocated number of bits which is higher than zero must be used to divide the spectrum into stopbands and passbands. The appropriate value for the threshold can be determined from bit allocation considerations such as above, from SNRSEG measurements, or listening tests. It was found that for 4.8 kbit/sec coded speech a threshold value of around 11 bits was appropriate.

Note that the bit allocation algorithm in this case is only used as a preliminary stage to implement a soft decision dividing the spectrum into stopbands and passbands. Once the division is decided, no further use is made of the number of allocated bits, since the minimization of P_I in 6.38 takes into account the relative importance of each band in the total reconstruction error. The role of the bit allocation algorithm is not therefore as crucial as in the previous coder structure.

Because of difficulties, already mentioned, in using the RMS_{m1} and RMS_{m2} parameters as side information, the side information parameter σ_{xk}^2 in 6.22 was chosen to be represented by the variance of the LPC residual in each band. This simplifies the algorithm. Furthermore the resulting codebook of gains does not depend on the bit allocation and pulse allocation algorithm as in the previous coder which provides a more robust system: the gains codebook, once obtained, can be used for a range of overall bit rates which was not possible in the previous coder.

Using the above threshold value of 11 bits it was found that the maximum allowable number of passbands is four. The percentage number of time blocks allocated a particular number of passbands is given below in table 6.T7.

number of passbands	1	2	3	4
% of time blocks	27.1	51.6	16.8	4.5

Table 6.T7: Percentage number of time blocks allocated a particular number of passbands

The above represent average values over 5 speakers (2 male and 3 female). Individual variations were found to be small.

6.10.2 Pulse Allocation

All passbands are allocated the same number of pulses. Furthermore these pulses are constrained to be at the same location (with respect to the beginning of the frame) across the bands. Each pulse set location and amplitudes are found iteratively as follows:

Assume that I-1 pulse sets have already been allocated. To find the shared location m_I and the optimum amplitudes g_{Ij} for each band j of the Ith pulse set proceed as follows: Let

$$P_{Ij} = \sum_{n=1}^{N} [e_{wj}{}^{I}(n)]^2 = \sum_{n=1}^{N} [X_{wj}(n) - \sum_{i=1}^{I} g_{ij} h_{wj}(n-m_i)]^2 \qquad 6.39$$

be the weighted error power for I pulses in band j $\{X_{wj}(n)\}$ are the input speech samples minus the memory from the previous frame and appropriately weighted, as in equation F5b of appendix F. $\{h_{wj}(n)\}$ are the impulse response samples of the modified filter as in F4 of appendix f. Let

$$P_I = \sum_{j=1}^{M} P_{Ij} \qquad 6.40$$

be the total weighted error power for I pulses. Here M is the total number of passbands.

The aim is to minimize P_I for each location m_I with respect to all pulse amplitudes g_{Ij}. This is the solution to the equations:

$$\frac{\partial P_I}{\partial g_{Ij}} = 0 \qquad j=1,2,\ldots M \qquad 6.41$$

from eq. 6.40 and 6.41

$$\sum_{k=1}^{M} \frac{\partial P_{Ik}}{\partial g_{Ij}} = 0 \qquad \begin{matrix} j=1,2,\ldots.M \\ k=1,2,\ldots.M \end{matrix} \qquad 6.42$$

equation 6.42 reduces to a set of M independent equations

$$\frac{\partial P_{Ij}}{\partial g_{Ij}} = 0 \qquad j=1,2,\ldots.M \qquad 6.43$$

since

$$\frac{\partial P_{Ik}}{\partial g_{Ij}} = 0 \qquad \text{for } j \neq k \qquad 6.44$$

Equation 6.39 can be written as

$$P_{jI} = \sum_{n=1}^{N} [e_{wj}{}^{I-1}(n) - g_{Ij}h_{wj}(n-m_I)]^2 \qquad 6.46$$

where

$$e_{wj}{}^{I-1}(n) = X_{wj}(n) - \sum_{i=1}^{I-1} g_{ij}h_{wj}(n-m_i) \qquad 6.46$$

can be considered to be the error at the I-1 stage. The solution of 6.43 using the expression in 6.45 for P_{jI} can be shown to be (appendix F)

$$g_{Ij} = \frac{\sum_{n=1}^{N} e_w^{I-1}(n) h_w(n-m_I)}{\sum_{n=1}^{N} [h_w(n-m_I)]^2} \qquad 6.47$$

or, through autocorrelation approximations

$$g_{Ij} = \frac{R_j e^{I-1} h\ (m_I)}{R_j hh(0)} \qquad 6.48$$

where the subscript w is implied.

The minimum power for location m_I can be shown to be (appendix F)

$$P_{jI}^{min} = P_{jI-1} - g_{Ij}^{2} R_j hh(0) \qquad 6.49$$

also, using 6.40

$$P_I^{min} = \sum_{j=1}^{M} [P_{jI-1} - g_{Ij}^{2} R_j hh(0)] \qquad 6.50$$

or

$$P_I^{min} = P_{I-1} - \sum_{j=1}^{M} g_{Ij}^{2}\ R_j hh(0) \qquad 6.51a$$

$$P_I^{min} = P_{I-1} - \sum_{j=1}^{M} \frac{R_j^{2}\ e^{I-1} h(m_I)}{R_j hh(0)} \qquad 6.51b$$

The location m_I is chosen (through exhaustive search over all m_i) which maximizes

$$\sum_{j=1}^{M} \frac{R_j{}^2 e^{I-1} h(m_I)}{R_j hh(0)} \qquad 6.52$$

The term $R_j e^I h(m_i)$, whose calculation is necessary in order to find the location and amplitude of the next pulse I+1 can be evaluated from the recursion:

$$R_j e^I h(m_i) = R_j e^{I-1} h(m_i) - g_{Ij} R_j hh(m_i - m_I) \qquad 6.53$$

with g_{Ij} given by 6.48.

6.10.3 Pulse locations and amplitudes

Let the number of passbands be M and the number of pulse sets allocated be K. The frame size is N(=16). Assuming 3 bit quantization for the amplitudes, the total number of bits required for the amplitudes is given by:

$$MxKx3 \quad \text{bits} \qquad 6.54$$

Using enumerative coding, the number of bits required for the locations is given by:

$$\log_2 \frac{N!}{K!(N-K)!} \quad \text{bits} \qquad 6.55$$

and numerically, rounded up to the nearest integer:

no. of pulses/band(k)	bits
1,15	4
2,14	7
3,13	10
4,12	11
5,11	13
6,10	13
7,9	14
8	14

Table 6.T8: Number of bits required for the encoding of the pulse locations.

The total bit requirement for K pulse sets can be found using (from 6.54 and 6.55)

$$B = 3.M.K + \log_2 \frac{N!}{K!(N-K!)} \quad \text{bits} \qquad 6.56$$

rounded up to the nearest integer.

Alternatively, knowing M and B, K can be found from 6.56. Some values are given in table 6.T9.

number of	Total bit rate (B)				
pulses (K)	1	2	3	4	number of passbands (M)
1	7	10	13	16	
2	13	19	25	31	
3	19	28	37	46	
4	23	35	47	59	
5	28	43	58		
6	31	49			
7	35	56			
8	38	62			
9	41				
10	43				
11	46				
12	47				
13	49				
14	49				
15	49				
16	48				

Table 6.T9: Number of bits required for the encoding of both the amplitudes and locations of the pulses

To maintain a bit rate not exceeding 4.8 kb/s, with 30 bits per 32 msec = 937.5 bits/sec for the LPC filters and (5 + 9) bits per 16 msec = 875 bits for the gains, giving a total of 1812.5 bits/sec for the side information, the pulse locations and amplitudes must be coded with no more than (4.800-1812.5) = 2987.5 bits or 47.8 bits/frame for a 16 msec frame. Therefore, the values in circles in table 6.T9 should be used. In the simulation, the values in squares were used for M = 1,2 which increase the bit rate by 12.5 and 75 bits/sec respecively. Therefore the maximum bit rate for this coder is 4.875 kbits/sec.

The coder was run in simulations using the above bit rates. The same codebooks were used for the LPC filters as in the previous coder. The gains side information was replaced by the rms of the LPC residuals and a new codebook was trained for these data. Empty bands were injected with white noise as before.

Although there was a noticeable improvement over the previous coder structure, particularly in terms of "burbling" sounds which were now eliminated completely, some of the background clicks and "knocking" sound still persisted. To identify the source of these distortions two simulations were performed. In the first case, the passbands were left uncoded i.e. the subband signals for the passbands were passed through, from the QMF analysis directly to the QMF synthesis without any processing, whilst the stopbands were injected with appropriately scaled noise as before (white noise was used to drive the corresponding LPC filters). In the second case the passbands were coded for 4.8 kbs/sec coding whereas the "stopbands" were passed through from the output of the QMF analysis into the QMF synthesis (i.e. the LPC residual signals were <u>not</u> replaced by white noise). Through these simulations it was found that the distortion was coming from the stopbands rather than the passbands.

6.10.4 Stopband regeneration

To improve the stopband excitation used above, a regeneration approach was used similar to the one that is used in certain RELP coders [7]. To understand the regeneration procedure a brief review of the QMF analysis-synthesis operations is necessary:

At the output of the QMF analysis filterbank 8 different signals are obtained. Each one of these signals occupies a bandwidth that extends from zero to 500 Hz. In effect, all subbands are frequency shifted to the baseband frequency range. In addition, the upper band generated at the output of each stage of the QMF-tree gets frequency inverted in the process. This has the effect that every other frequency band (starting with the second) not only gets translated to the baseband but it is also frequency inverted.

For the passbands, the pulses model the envelope flattened subband signals. Each of these excitation signals is passed through the corresponding LPC synthesis filter and fed through the QMF synthesis tree which, in effect, translates each subband signal to its appropriate frequency range, frequency inverting the bands to their original orientations at each stage and, finally sums the outputs to produce the full-band signal.

The QMF synthesis tree accepts 8 lowpass inputs (0-500 Hz), which are frequency translated, through the tree, into 8 contiguous, 500 Hz wide, bandpass slots to occupy the frequency range from 0-4000 Hz.

Consider the situation where only one passband is allowed, the one that would occupy the bandpass slot 0-500 Hz in the fullband signal at the output of the QMF tree. Further assume that the excitation signal (i.e. the pulses generated through the algorithm) from this band is duplicated for the other 7 inputs of the QMF tree. In addition, these other excitation signals are amplitude scaled such that their variances match that of the LPC residual signal in the corresponding band and, finally, are passed through the corresponding LPC synthesis filters.

The last two operations guarantee that the envelope of the synthesized fullband signal is a close match to the envelope of the uncoded fullband signal. As far as the envelope-flattened subband signals are concerned, the effect of duplicating the excitation signal for the other 7 inputs to the tree would be identical to that of spectral folding in figure 3.2-59b (due to the frequency inversion in the QMF synthesis). If every other excitation subband signal (starting with the second one) is in addition frequency inverted before scaling and passed through the LPC and QMF synthesis then the effect would be identical to spectral translation as in figure 3.2-59c. The full regeneration would then be equivalent to the RELP regeneration in figure 3.2-60 for spectral folding and figure 3.2-61 for spectral translation.

In practice, more than one passband is selected. The total excitation which is the sum of the excitations from all the passbands can then be used to drive the filters, correspnding to the stopbands. The rest of the operations are identical to the above procedure. Care was taken before adding the excitations together to revert any frequency inverted signals back to their original orientations. Spectral translation was found to be perceptually more acceptable than spectral folding which results in (barely) audible tones. The performance improves further by subtracting the short term dc level from the composite excitation signal prior to synthesis.

The regeneration technique improved the speech quality dramatically, considering the fact that no extra bits were required for the improvement. The "click" and "knocking" sounds had disappeared completely.

6.10.5 A modification to the coding of the LPC parameters

Although the obtained speech quality was generally good, for certain (female) speakers certain speech sounds wee degraded by a "thumping" sound which although it occurred rarely it was perceptually annoying. The origin of this distortion was identified to lie within the quantized LPC filters since the degradation was absent when the filters were used unquantized in the simulations. The filters' bit rate was raised accordingly by using a 9 bit codebook for the first band and 8 bit codebooks for the rest of the bands. This increased the bit rate for the LPC filters by about 100%. Surprisingly, little improvement over the above distortion was observed. This was unexpected since with unquantized filters the distortion was completely absent. After careful consideration it was decided that the distortion criterion used in vector encoding the LAR's must have been an inappropriate measure, at least for this case of subband filters. This would introduce a random element in the selection of codewords both in the training phase as well as in the actual quantization coding, which could explain why doubling the rate introduced a very small improvement.

The function of the LPC filter is to reduce the variance of the signal to be encoded thus reducing the error variance. This suggests that, the signal to residual error variance ratio could serve as a useful distortion criterion. Alternatively, the ratio formed by dividing the residual variance obtained with quantized filters by the residual variance obtained with unquantized filters could serve as an appropriate criterion. This is the same as the likelihood ratio δ/α

given by 3.2-125. In order to assess the applicability of δ/α as a useful distortion criterion a threshold value TH was set up and compared, for each frame and for each band during coding, with the function

$$F = 10\log_{10} \frac{\delta}{\alpha} \text{ dB} \qquad 6.57$$

If F was greater than TH then the unquantized filters were used for coding whereas if F was less than TH the quantized filter (where quantization was performed with the m.s.e. criterion on the LAR as before) was used instead. The number of unquantized filters used as a percentage of the total (note that percentages from different bands were pooled together) was measured for several values of TH. The SNRSEG measure between coded and original signals was also monitored. The results are shown in table 6.T10 and in figure 6.29.

Threshold value TH (dB)	∞	7	5	3	2	−∞
Unquantized filters as a percentage of total (%)	0.0%	2.81	5.99	13.52	24.20	100
SNRSEG (dB)	8.33	8.95	9.17	9.55	9.51	9.71

Table 6.T10: Segmental SNR as a function of the percentage of the unquantized frames

It can be seen from figure 6.29 that the SNRSEG improvement saturated with as little as 15% of the quantized filters been replaced with unquantized ones. Perceptually, even with the threshold at 5 dB (6% unquantized filters) the performance was nearly as good as with the threshold at (-∞) (100% unquantized filters) and with the threshold at 3 dB (13.5%) it was only through careful listening with headphones that any difference could be observed between this value and a threshold at -∞. The good correlation of the threshold value with objective and subjective quality demonstrated that the residual error power was a good distortion criterion for encoding the LAR.

In order to incorporate this knowledge into the algorithm, in the vector quantization procedure that vector was chosen for transmission which minimized the residual power in that band instead of the m.s.e. between the LAR. Note that the codebook used was the original one obtained using a m.s.e. criterion on the LAR. The resulting speech quality was nearly as good as with unquantized filters. No "thumping" distortion was evident. It is expected that even better quality could be obtained if the codebooks were trained using a distortion criterion related to the likelihood ratio as opposed to the m.s.e. on the LAR.

6.10.6 Speech quality obtained from the final algorithm

The resulting algorithm is shown in figures 6.30a (encoder) and 6.30b (decoder).

SNRSEG values obtained with 4 typical sentences (2 male and 2 female) are given in table 6.T11 in comparison with similar results from the fullband multipulse algorithm [6], both coded at 4.8 kb/s

	DANCE(F)	SPEECH (M)	MICHAEL (M)	STREAM (F)
multipulse	7.05	6.09	7.77	7.27
proposed algorithm	9.37	6.87	9.41	9.22

DANCE: THEY DANCED IN EXCITEMENT AROUND THE FIRE

SPEECH: INDUSTRIAL SHARES WERE MOSTLY A TRIFFLE HIGHER

MICKEL: THERE WAS AN OLD MAN CALLED MICHAEL FINNEGAN

STREAM: AT THE SIDE OF THE ROCK A SMALL STREAM FLOWED INTO THE RIVER.

Table 6.T11: Segmental SNR results obtained with 4 different speakers

The perceptual quality of the proposed algorithm was far superior compared to that of the corresponding multipulse algorithm, especially for female voices. The coded speech had a smooth and "pleasant" quality.

6.10.7 Coder Complexity

In its present form the proposed coder scheme requires about 160 mult/add operations per full band sample (1/8 msec). This is mostly due to the QMF analysis-synthesis (96 mult/add operations per sample). The large complexity of the QMF structure arises from the fact that 32 tap filters were used for each of the three stages of QMF analysis-synthesis.

It is well known that the same performance can be obtained using less complex QMF structures (e.g. Parallel QMF) and the complexity of the QMF can be reduced to less than half of its present vale. Therefore, the complexity of the proposed coder can be made comparable (if not less) to that of the simplified multipulse coder of [6]. The detailed calculation of the coder complexity as well as that of [6] is given in appendix H.

6.10.8 Further work

The bit rate of the algorithm can be reduced further if the remaining redundancies are exploited. The gains vector as well as the overall gain parameter can be differentially encoded to take advantage of frame to frame correlations. The same can be done for the LPC filters. In addition, for the LPC filters, the correlation between the filters from different bands (for the same time frame) can be exploited. One way to do this is to apply a KL transform to the vector formed by concentrating all the filter vectors together, prior to quantizing each filter individually.

6.10.8.1 Vector Quantization of the pulse amplitudes

Equation 6.45 gives the m.s.e. for unquantized pulse amplitudes g_{Ij}. Let the quantized pulse amplitudes be given by $\hat{g}_{Ij}$. Equation 6.45 can be written as:

$$P_{jI} = \sum_{n=1}^{N} [e_{wj}^{I-1}(n) - \hat{g}_{Ij}h_{wj}(n-m_I)]^2 \qquad 6.58$$

were the term $e_{wj}^{I-1}(n)$ now includes the error due to the quantization of the pulse amplitudes $\{g_{ij}$ $i = 1,2,\ldots I-1\}$. Expanding 6.58 gives

$$P_{jI} = P_{jI-1} - 2\hat{g}_{Ij}\sum_{n=1}^{N} e_{wj}^{I-1}(n)h_{wj}(n-m_I) +$$

$$\hat{g}_{Ij}^{2} \sum_{n=1}^{N} h_{wj}(n-m_I)^2 \qquad 6.59$$

Using 6.47:

$$P_{jI} = P_{jI-1} - (2g_{Ij}\hat{g}_{Ij} - \hat{g}_{Ij}^{2}) \sum_{n=1}^{N} h_{wj}^{2}(n-m_I) \qquad 6.60$$

where g_{Ij} is the optimum unquantized amplitude.

Let

$$\hat{g}_{Ij} = g_{Ij} + q_{Ij} \qquad 6.61$$

where q_{Ij} is the quantization error in quantizing g_{Ij}. The term in brackets in 6.60 can then be expressed as

$$2g_{Ij}\hat{g}_{Ij} - \hat{g}_{Ij}^{2} = g_{Ij}^{2} - q_{Ij}^{2} \qquad 6.62$$

Therefore 6.60 is given by

$$P_{jI} = P_{jI-1} - (g_{Ij}{}^2 - q_{Ij}{}^2) \sum_{n=1}^{N} h_{wj}{}^2(n-m_I) \qquad 6.63$$

and the overall (weighted) error power is minimized when the term

$$Q_I = \sum_{j=1}^{M} q_{Ij}{}^2 \sum_{n=1}^{N} h_{wj}{}^2(n-m_I) \qquad 6.64$$

is minimized.

Let G_j be the jth (quantized) component of the gains vector. Expanding 6.64 using 6.61:

$$Q_I = \sum_{j=1}^{M} [\frac{g_{Ij}}{G_j} - \frac{\hat{g}_{Ij}}{G_j}]^2 G_j{}^2 \sum_{n=1}^{N} h_{wj}{}^2 (n-m_I) \qquad 6.65$$

The minimization of Q_I is equivalent to a nearest neighbour search between an input vector given by $\{g_{Ij}/G_j,\ j=1,2,....M\}$ a codeword $\{\hat{g}_{Ij}/G_j,\ j=1,2,....M\}$ and a weighted mean square error criterion with weights

$$W_j = G_j{}^2 \sum_{N=1}^{N} h_{wj}{}^2(n-m_I) \qquad 6.66$$

The factor G_j is used so that the space in which the input vectors are situated is made as small as possible thus increasing the quantization efficiency.

Note that the input vector dimension M can take several values (M = 1,2,3 or 4 in the above algorithm). This will require the design of a separate codebook for different values of M (The number of levels in each codebook will also be different).

Vector quantization of the pulse amplitudes should lead to a reduction of the pulse amplitudes bit rate which is a significant term in the overall bit rate. Equation 6.65 describes the required codebook design completely.

6.10.9 Note on Publication

A paper entitled "A new approach to low bit rate coding" has been presented at the Fifth International Conference on Digital Processing of Signals in Communications that was held in Loughborough University of Technology in September 1988. This paper was written in co-authorship with Professor C.S. Xydeas and covers the work presented in section 6.10 of this chapter.

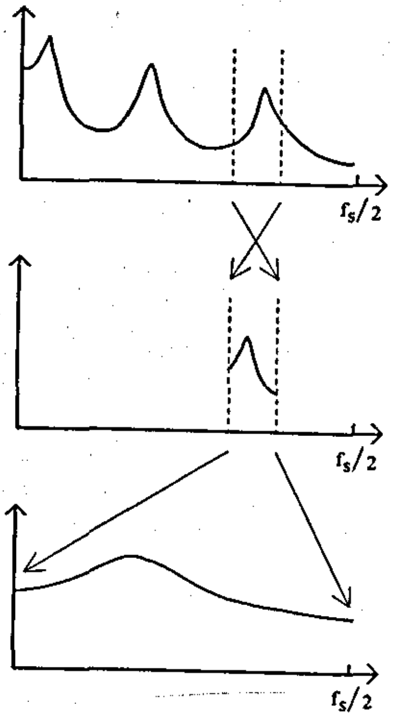


Figure 6.1 A schematic diagram of the frequency domain operations to derive a bandpass envelope from the fullband envelope. (The effects of the QMF amplitude responce are neglected.)

Fig. 6.2 First Band Data.

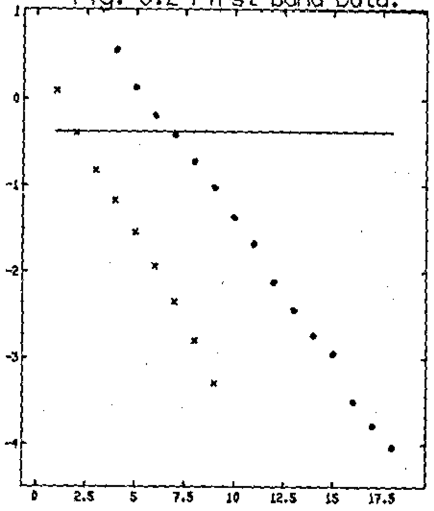


Fig. 6.3 Second Band Data.

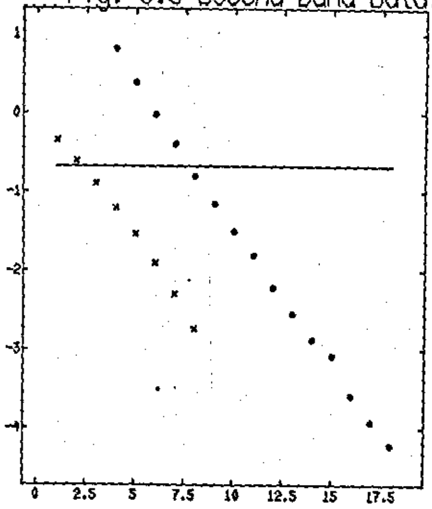


Fig. 6.4 Third Band Data.

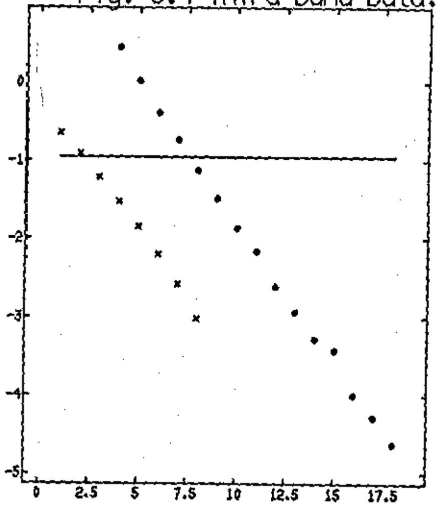


Fig. 6.5 Fourth band Data.

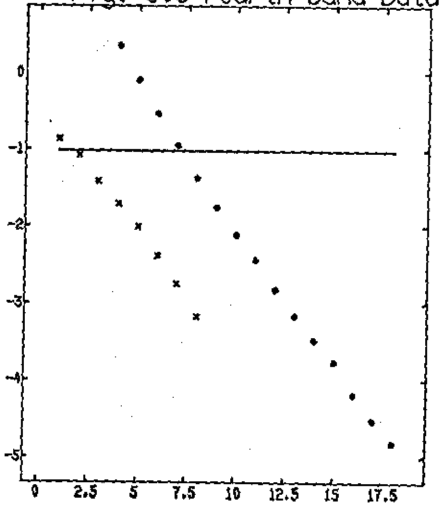


Fig. 6.6 Fifth Band Data.

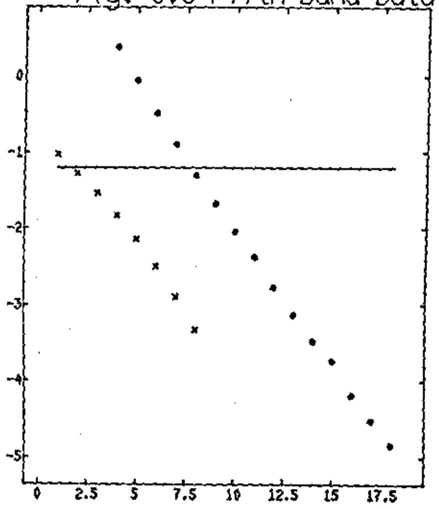


Fig. 6.7 Sixth Band Data.

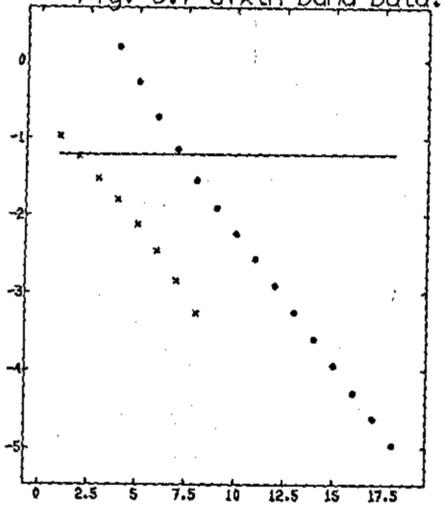
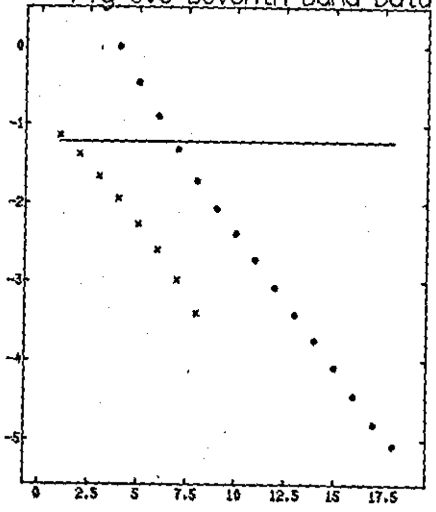


Fig 6.8 Seventh Band Data.



Figures 6.2 – 6.8 A comparative study of the logarithm of the mean square LAR distortion (ordinate) as a function of the number of bits (abscissa). Crosses : vector quantizer, Circles : scalar quantizer, Straight line : vector quantizer employing inferred codebooks.

Fig. 6.9 First Band Data.

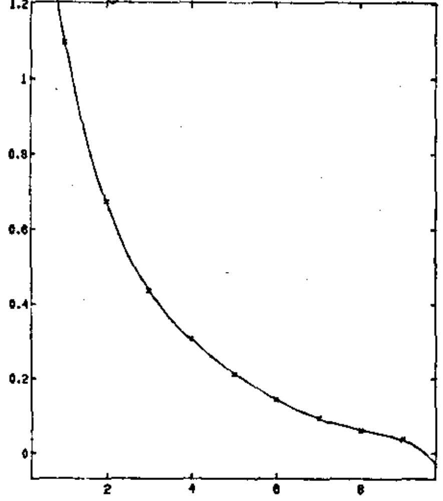


Fig. 6.10 Second Band Data.

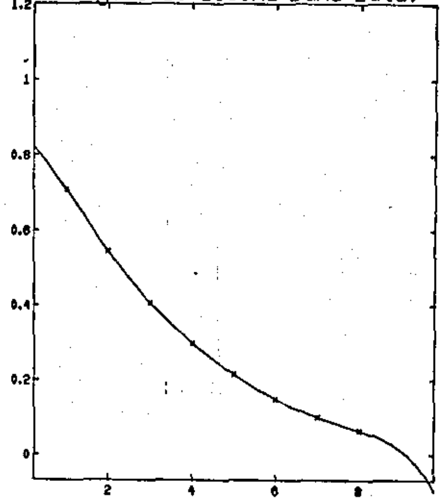


Fig. 6.11 Third Band Data.

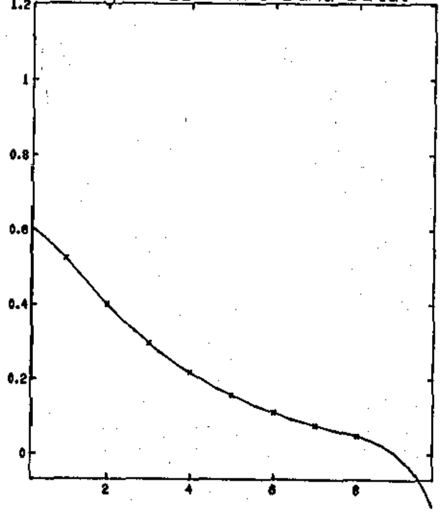


Fig. 6.12 Fourth band Data.

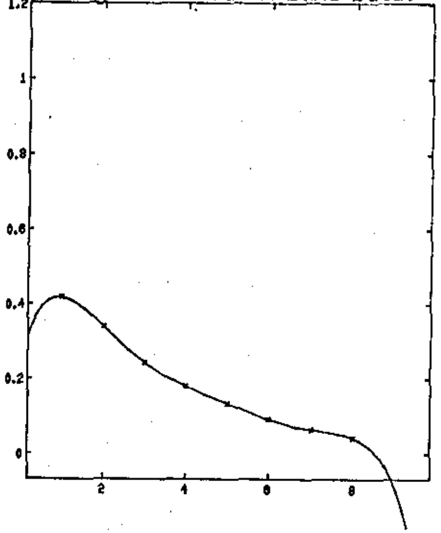


Fig. 6.13 Fifth Band Data.

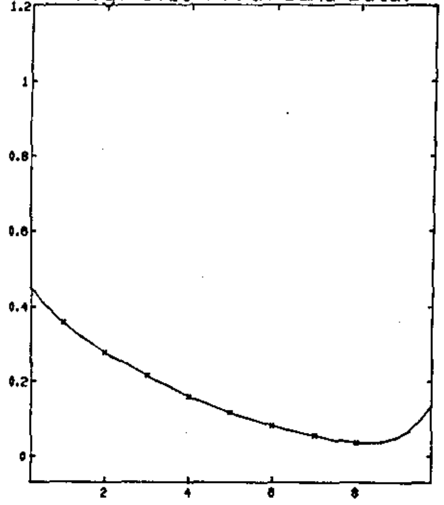


Fig. 6.14 Sixth Band Data.

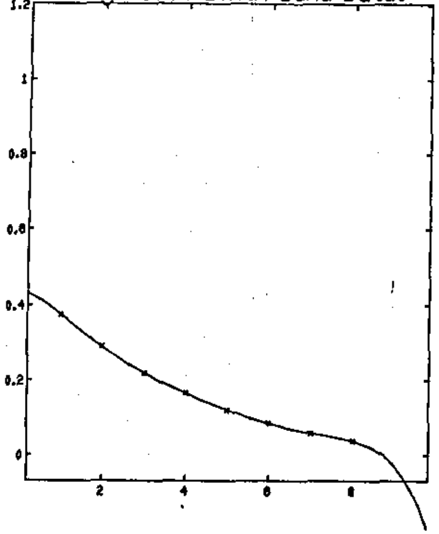
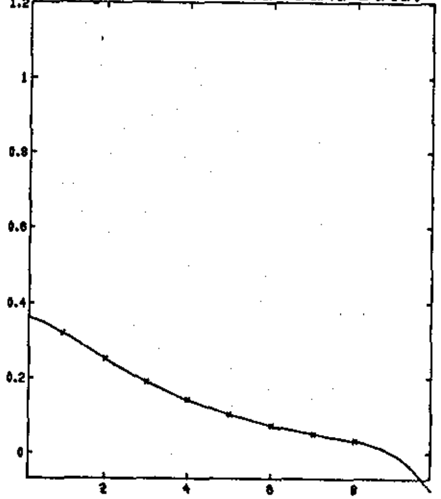


Fig. 6.15 Seventh Band Data.



Figures 6.9 – 6.15 Mean square LAR distortion (ordinate) as a function of the number of bits (abscissa), for the seven LPC filter codebooks (crosses). The sixth order polynomials fitted to these data are also shown (solid lines).

Figure 6.16

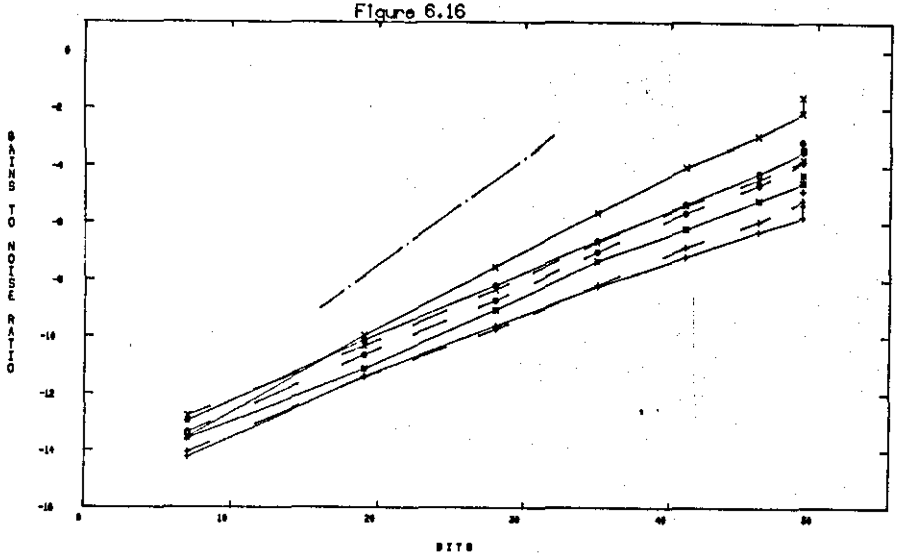


Figure 6.17

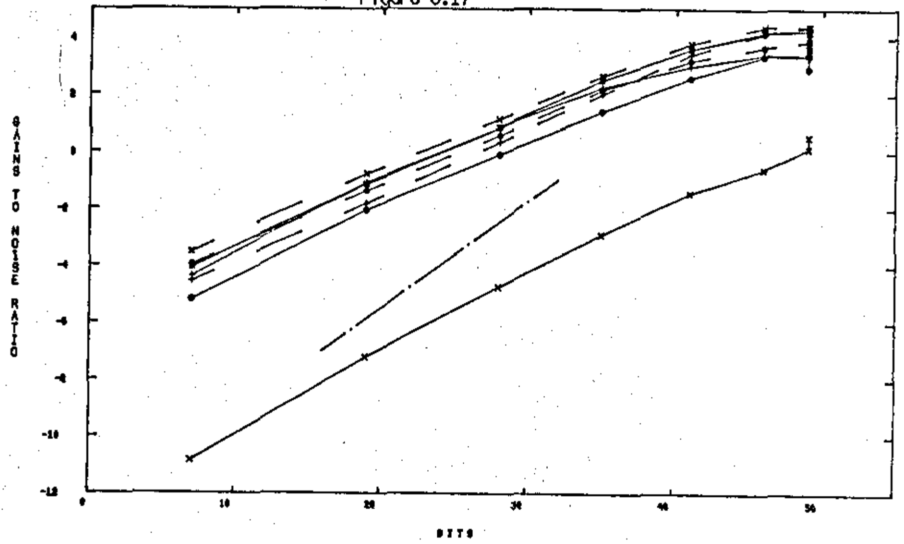


Figure 6.18

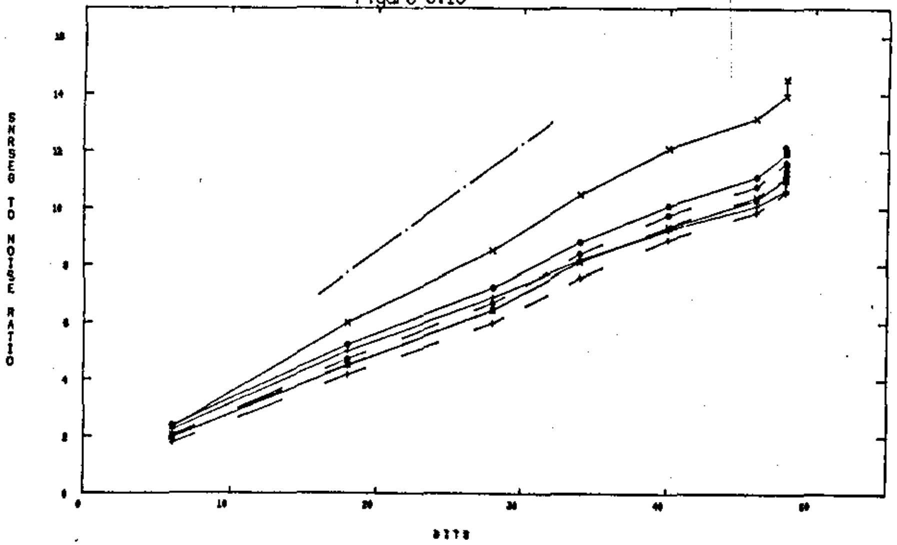


Figure 6.16 Segmental gain to noise ratio. RMSm1 is used to represent the gains. The 6 dB/bit line is also shown for comparison.

Figure 6.17 Segmental gain to noise ratio. RMSm2 is used to represent the gains. The 6 dB/bit line is also shown for comparison.

Figure 6.18 Segmental signal to noise ratio.

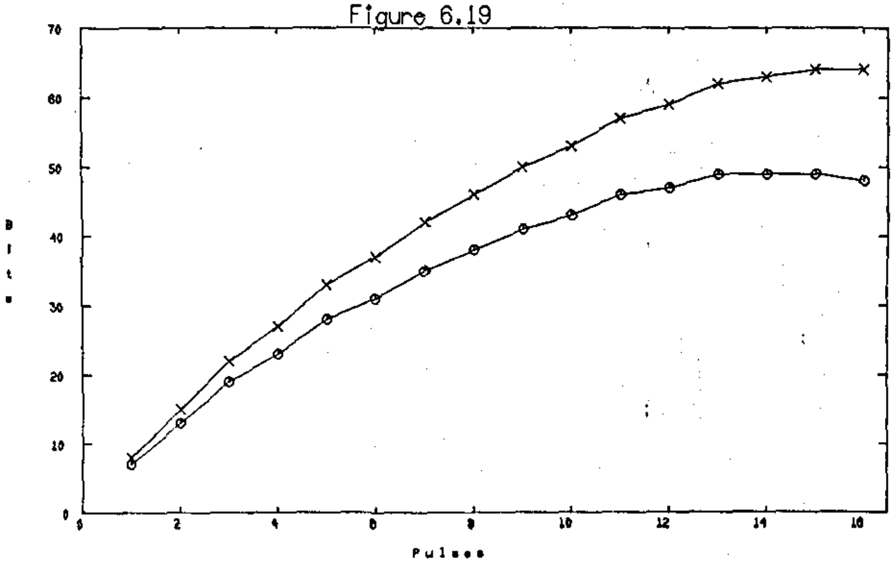
Solid curve :

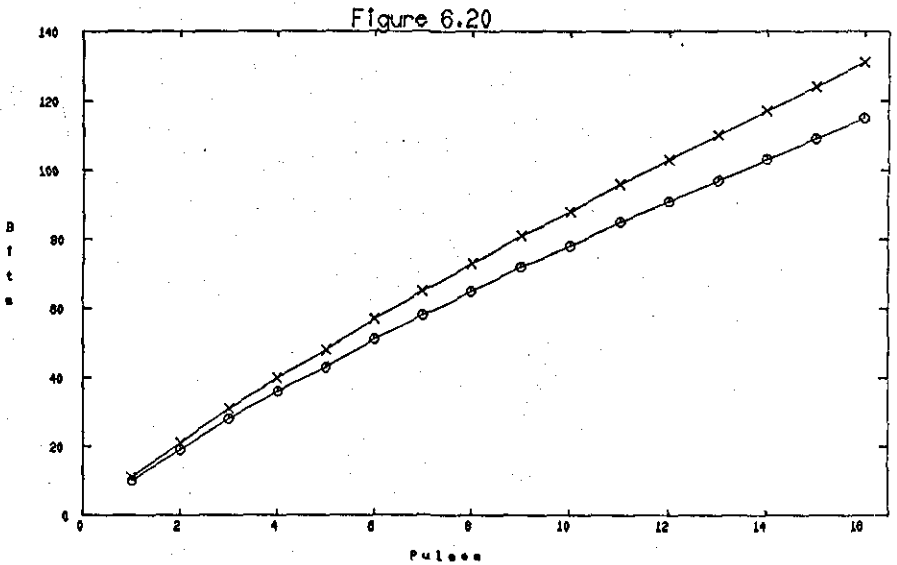
x : first band, o : second band, + : third band, * : fourth band.

Dashed curve :

x : fifth band, o : sixth band, + : seventh band.

Dot–dashed curve : 6 dB/bit line.





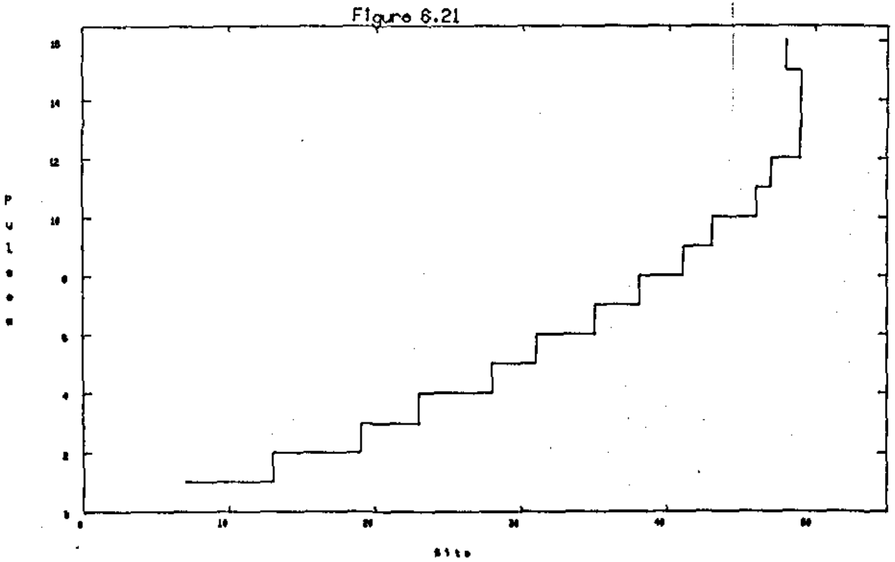


Figure 6.19 Bit rate as a function of the number of pulses for a frame size of 16 samples. Circles : three bit amplitude quantization. Crosses : four bit amplitude quantization.

Figure 6.20 Bit rate as a function of the number of pulses for a frame size of 128 samples. Circles : three bit amplitude quantization. Crosses : four bit amplitude quantization.

Figure 6.21 Number of pulses as a function of bit rate (3 bits per pulse) subject to an integer pulse constraint. The frame size is 16 samples.

Table 6.T3 Bit requirements for a 16 msec frame. Number of samples in the frame=16, number of bits per pulse amplitude=3

No of pulses	Bits/frame	Bits/sec	Bits/pulse
1	7	437.5	7.0
2	13	812.5	6.5
3	19	1187.5	6.3
4	23	1437.5	5.8
5	28	1750.0	5.6
6	31	1937.5	5.2
7	35	2187.5	5.0
8	38	2375.0	4.8
9	41	2562.5	4.6
10	43	2687.5	4.3
11	46	2875.0	4.2
12	47	2937.5	3.9
13	49	3062.5	3.8
14	49	3062.5	3.5
15	49	3062.5	3.3
16	48	3000.0	3.0

Table 6.T4 Bit requirements for a 16 msec frame. Number of samples in the frame=16, number of bits per pulse amplitude=4

No of pulses	Bits/frame	Bits/sec	Bits/pulse
1	8	500.0	8.0
2	15	937.5	7.5
3	22	1375.0	7.3
4	27	1687.5	6.8
5	33	2062.5	6.6
6	37	2312.5	6.2
7	42	2625.0	6.0
8	46	2875.0	5.8
9	50	3125.0	5.6
10	53	3312.5	5.3
11	57	3562.5	5.2
12	59	3687.5	4.9
13	62	3875.0	4.8
14	63	3937.5	4.5
15	64	4000.0	4.3
16	64	4000.0	4.0

Table 6.T5 Bit requirements for a 16 msec frame. Number of samples in the frame=128, number of bits per pulse amplitude=3

No of pulses	Bits/frame	Bits/sec	Bits/pulse
1	10	625.0	10.0
2	19	1187.5	9.5
3	28	1750.0	9.3
4	36	2250.0	9.0
5	43	2687.5	8.6
6	51	3187.5	8.5
7	58	3625.0	8.3
8	65	4062.5	8.1
9	72	4500.0	8.0
10	78	4875.0	7.8
11	85	5312.5	7.7
12	91	5687.5	7.6
13	97	6062.5	7.5
14	103	6437.5	7.4
15	109	6812.5	7.3
16	115	7187.5	7.2

Table 6.T6 Bit requirements for a 16 msec frame. Number of samples in the frame=128, number of bits per pulse amplitude=4

No of pulses	Bits/frame	Bits/sec	Bits/pulse
1	11	687.5	11.0
2	21	1312.5	10.5
3	31	1937.5	10.3
4	40	2500.0	10.0
5	48	3000.0	9.6
6	57	3562.5	9.5
7	65	4062.5	9.3
8	73	4562.5	9.1
9	81	5062.5	9.0
10	88	5500.0	8.8
11	96	6000.0	8.7
12	103	6437.5	8.6
13	110	6875.0	8.5
14	117	7312.5	8.4
15	124	7750.0	8.3
16	131	8187.5	8.2

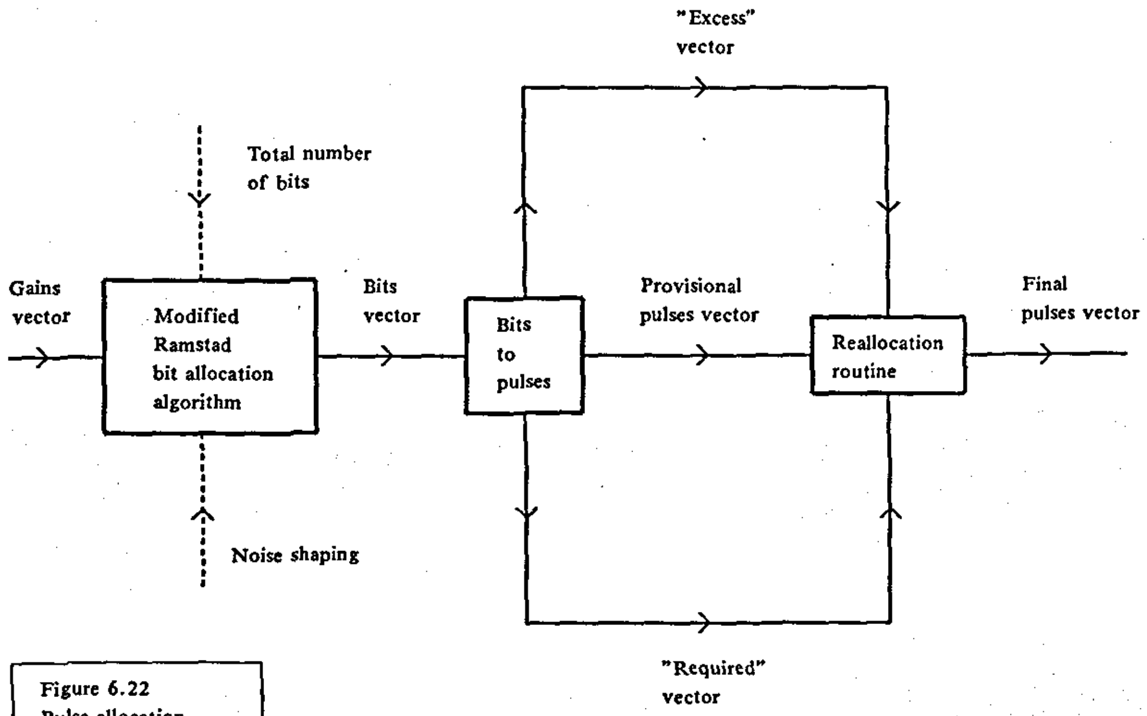


Figure 6.22
Pulse allocation
routine - Flowchart

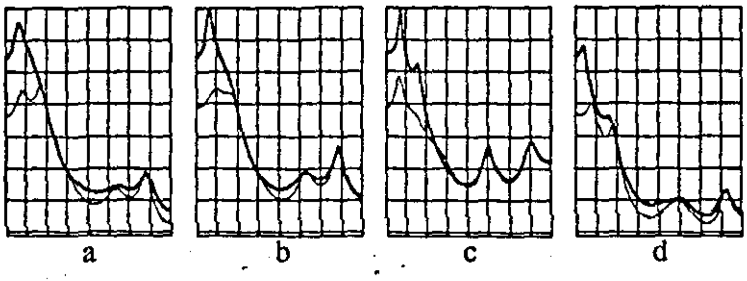


Figure 6.23 a-d LPC speech (thick line) and noise (thin line) spectra of successive speech frames. The Gains side information was left uncoded.

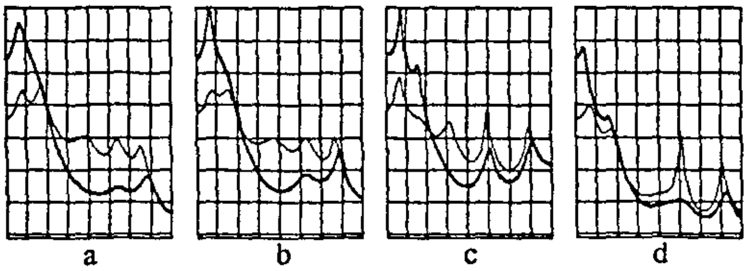


Figure 6.24 a-d LPC speech (thick line) and noise (thin line) spectra of successive speech frames. The Gains side information was coded using a m.s.e. criterion on the components of G_k.

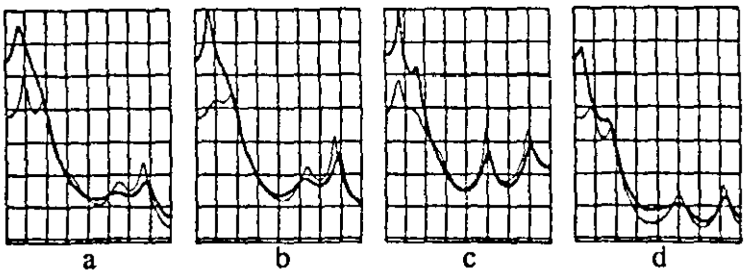


Figure 6.25 a-d LPC speech (thick line) and noise (thin line) spectra of successive speech frames. The Gains side information was coded by logarithmically compressing the components of G_k prior to quantization.

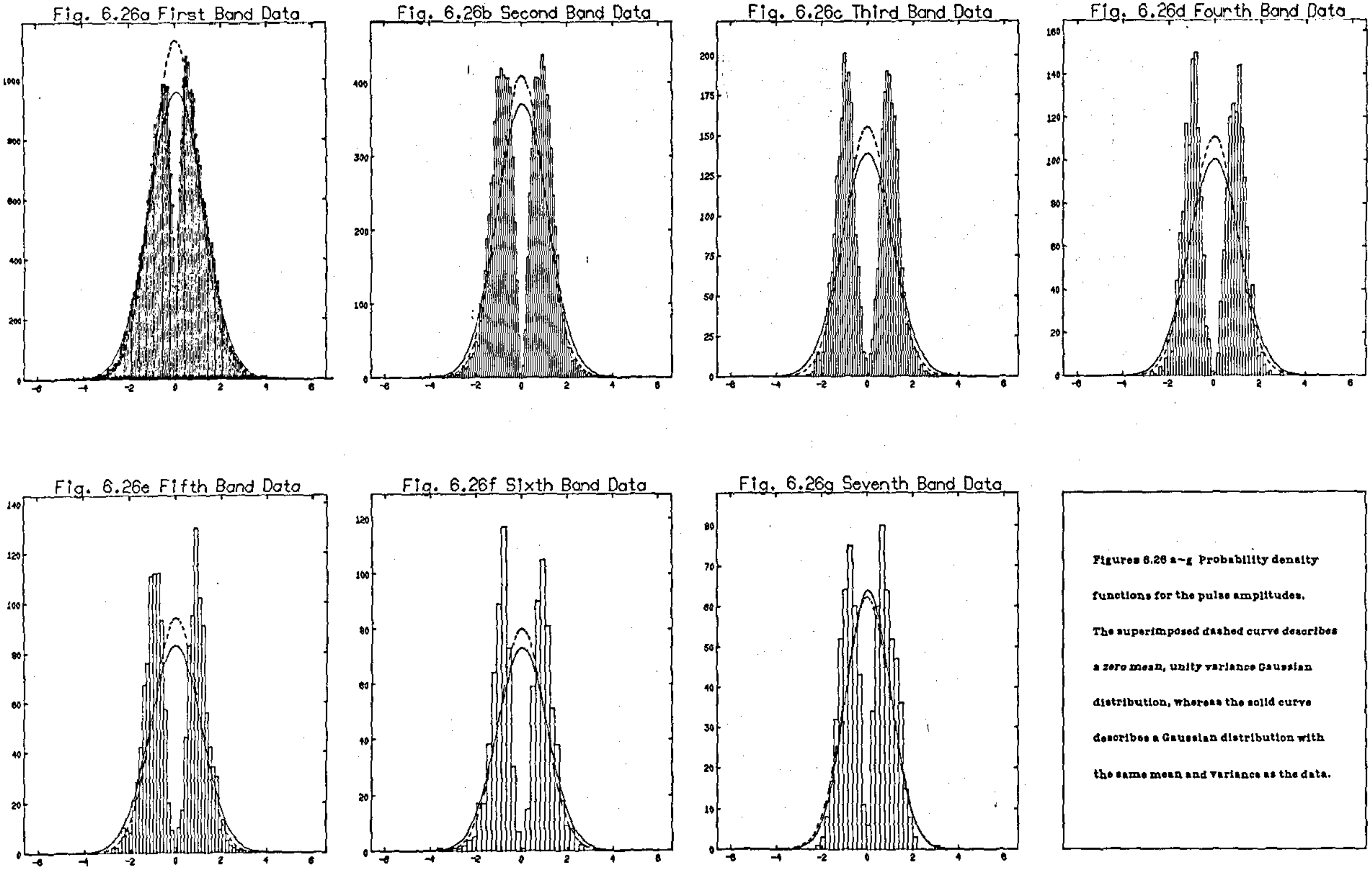


Fig. 6.26a First Band Data

Fig. 6.26b Second Band Data

Fig. 6.26c Third Band Data

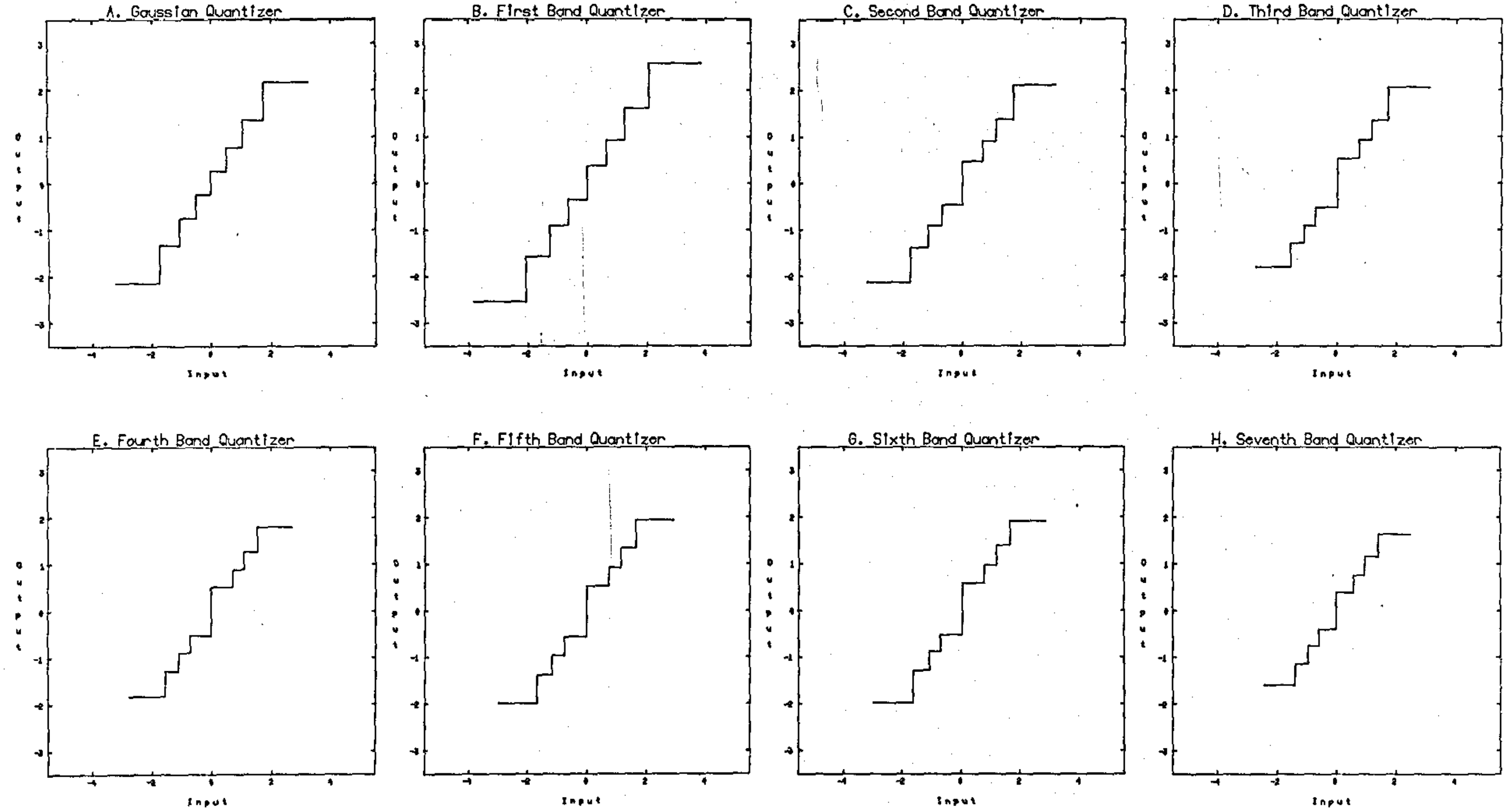
Fig. 6.26d Fourth Band Data

Fig. 6.26e Fifth Band Data

Fig. 6.26f Sixth Band Data

Fig. 6.26g Seventh Band Data

Figures 6.26 a–g Probability density functions for the pulse amplitudes. The superimposed dashed curve describes a zero mean, unity variance Gaussian distribution, whereas the solid curve describes a Gaussian distribution with the same mean and variance as the data.



Figures 6.27 B–H Optimum (Max–Lloyd) quantizer characteristics for the seven bands covering the frequency range between 0–3.5 KHz. A Gaussian quantizer characteristic is also shown (figure 6.27 A).

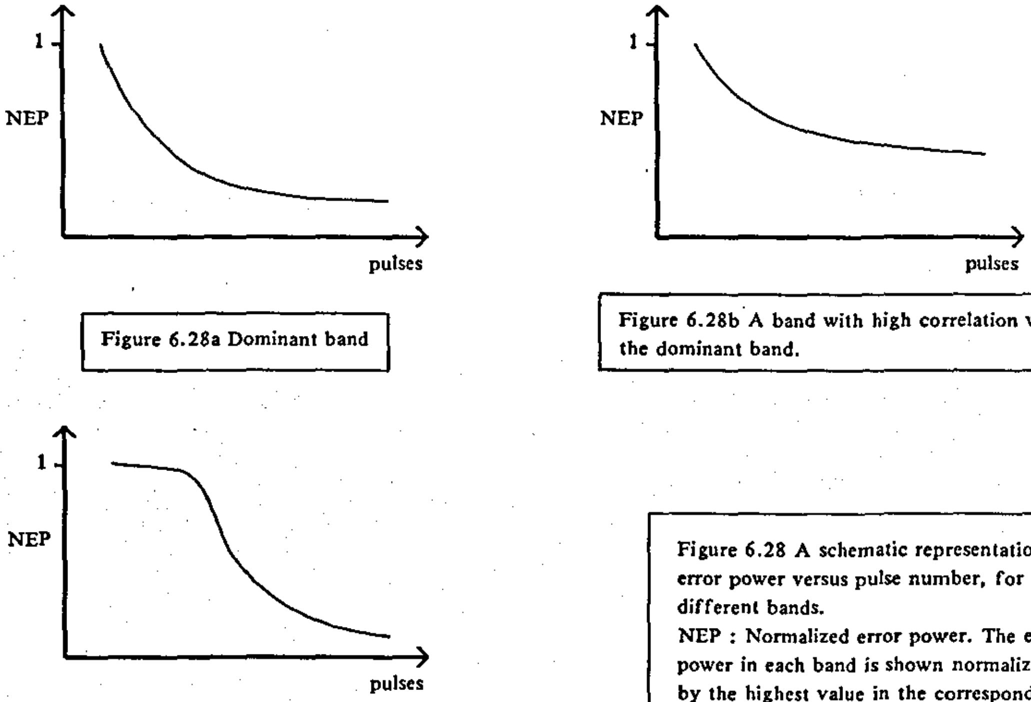


Figure 6.28a Dominant band

Figure 6.28b A band with high correlation with the dominant band.

Figure 6.28c A band with low correlation with the dominant band.

Figure 6.28 A schematic representation of error power versus pulse number, for three different bands.
NEP : Normalized error power. The error power in each band is shown normalized by the highest value in the corresponding band.

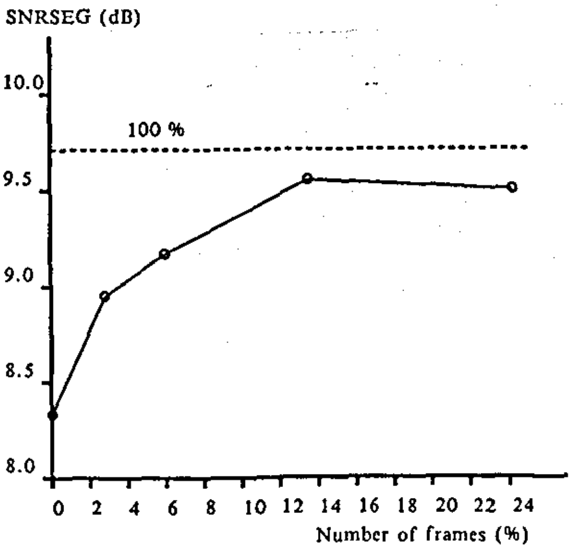


Figure 6.29 SNRSEG results as a function of the number of frames with unquantized LPC coefficients(expressed as a percentage of total.) The dashed line shows the SNRSEG value with no LPC coefficient quantization(100 %).

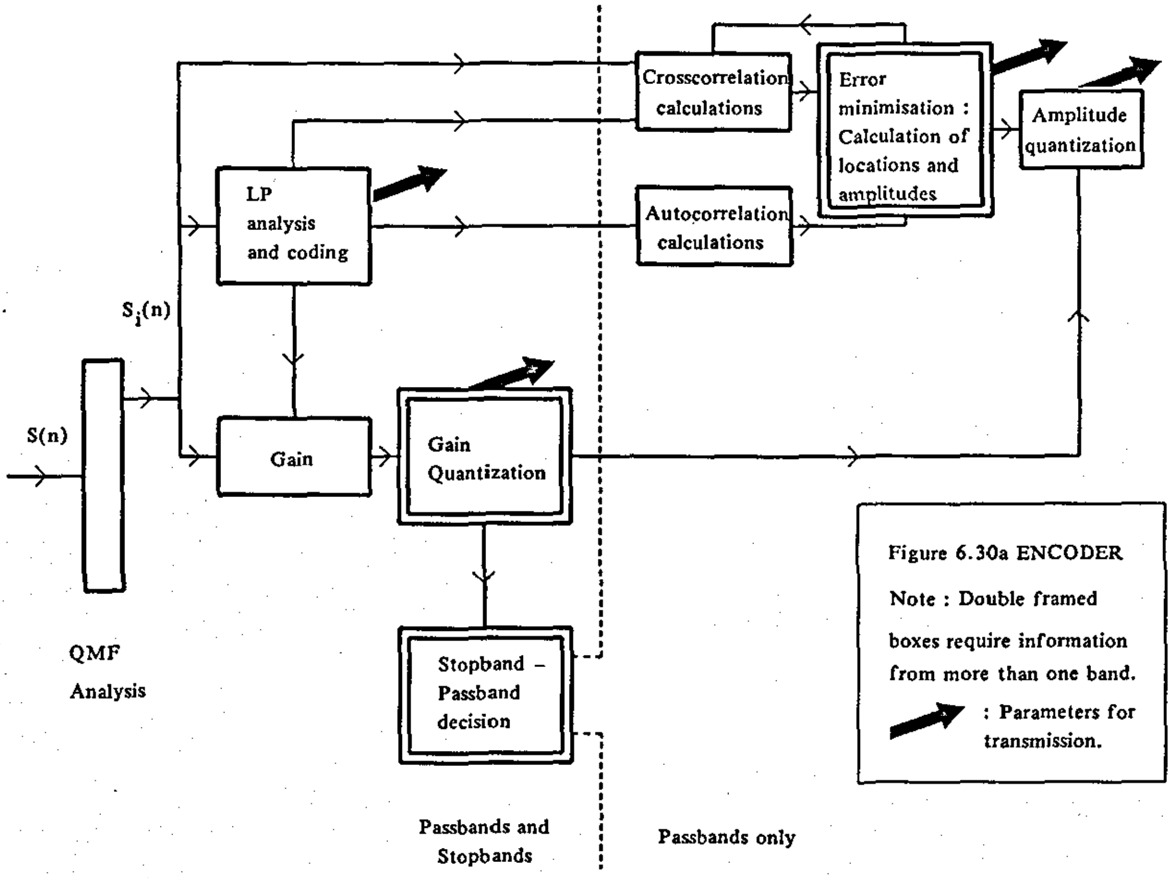

Crosscorrelation calculations
Error minimisation : Calculation of locations and amplitudes
Amplitude quantization
LP analysis and coding
Autocorrelation calculations
$S_i(n)$
S(n)
Gain
Gain Quantization
Stopband – Passband decision
QMF Analysis
Passbands and Stopbands
Passbands only

Figure 6.30a ENCODER

Note : Double framed boxes require information from more than one band.

: Parameters for transmission.

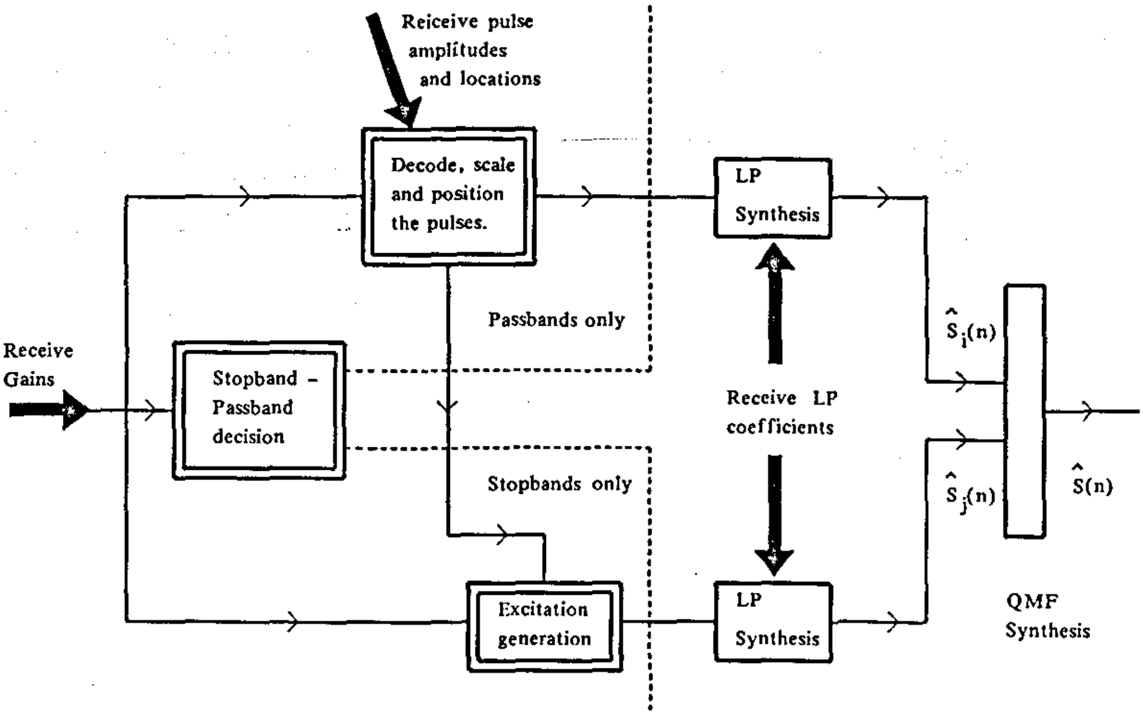


Figure 6.30b. DECODER

Note : Double framed boxes require information from more than one band.

1. R. Viswanathan, J. Makhoul "Quantization Properties of Transmission Parameters in Linear Predictive Systems", IEEE Trans. on Acoustics Speech and Signal Proc. Vol. ASSP-23, No 3, June 1975, pp309-321

2. J. Makhoul, S. Roucos, H. Gish "Vector Quantization in speech Coding" IEEE Proc. Vol. 73, No. 11, Nov. 1985, pp1551-1588.

3. NAG Fortran library manual, mark 11.

4. V. Ramamoorthy and N.S. Jayant "Enhancement of ADPCM speech by adaptive postfiltering" Bell system Tech. J. 63(8) pp1465-1475 Oct. 1984.

5. J.H. Chen and A. Gersho "Real-time vector APC speech coding at 4.8 kb/s with adaptive postfiltering" IEEE Proc. ICASSP 1987 pp2185-2188.

6. T. Araseki, K. Ozawa, S. Ono, K. Ochiai "Multipulse excited speech coder based on maximum crosscorrelation search algorithm" IEEE Proc. Global Telecommun. Conf.1983 pp794-798.

7. J. Makhoul, M. Berouti, "Predictive and Residual Encoding of Speech" J. Acoust. Soc. Am. pp1633-1640, 1979.

CHAPTER 7:

RECAPITULATION AND CONCLUSIONS -

DIRECTIONS FOR FUTURE RESEARCH

The emphasis in the present work is in the inegration of psychoacoustic knowledge and mathematically sound, analytically tractable algorithms, in the fields of speech quality assessment and coding.

The first contribution that this thesis has to offer is the extensive review of physiological and psychophysical phenomena, of relevance to speech research.

This comprehensive and functional description of relevant mechanisms - as is provided by this multi-disciplinary analysis - will hopefully serve as a valuable reference source to enhance further understanding and application of auditory perceptual concepts in speech research.

This review appears in chapter 2. However, applications of perceptual knowledge will be found throughout the thesis.This is also the case for chapter 3:Although a literature review of various rate compression algorithms, it frequently provides psychoacoustical reasoning for the resulting (perceptual) speech quality.

The application of perceptual knowledge is perhaps most explicitly madë in chapter 4, where a model is developed. This aims to simulate the way human subjects evaluate speech degradation resulting form rate compression. The model is comprehensive in that it does not only simulate the functions of the peripheral auditory system (PAS) but, in addition, the subsequent processing of auditory information by the condition centre in the brain.

The simulation of the PAS is achieved by processing the signal via the same transformations that are known to operate within the outer, middle and inner ear. The resultant representation of the auditory signals is not unlike that at the output of a running spectrum analyser. In the present case though, the notions of a frequency axis, the spectral resolution and spectral magnitudes are redefined to comply with their auditory equivalents.

The few attempts, that can be found in the literature, in constructing such models have traditionally concentrated in obtaining smooth, "envelope" representations, more in line with "place" theories in hearing. The approach in the present work was to incorporate such elements in the model to make it applicable to the measurement of time varying events. Such events are typical of speech sounds. The resulting algorithms combine elements from both "place" and "volley" theories. This complies with recent physiological and psychophysical results obtained with temporary variable sounds. However, the idea of preferential weighting of different time segments merits further investigation.

The mechanisms by which human observers classify and measure different types of distortion was also investigated. Multidimensional procedures were used to create a distortion space. The origin of this space is occupied by the uncoded speech signal. Different directions in the space are related to different types of distortion. Distortion axes are identified, such as the "hissing (PCM-type)"-axis and the "Roughness" axis. Coded speech signals are represented as points within this space. The distance of each point from the origin (Original signal) is a measure of its overall distortion. The projections of this distance onto the distortion specific axes decomposes the distortion into its particular types. This indicates how much distortion of each type is present in the coded speech segment in question. The distance between different coded files is related to the dissimilarity of distortion present in each one. The dimensionality of the space is a function of the number of distinct distortions that can be identified.

Various procedures are used to map the distortion space onto spaces of lower dimensionality, which facilitate visual assessment of the distortions. These procedures ensure that the initial distances in the higher dimensionality space are closely approximated in the lower dimensionality space.

Similar distortion spaces are obtained from the objective tests described above as well as from appropriately conducted subjective tests, proving the validity of the adopted approach.

The inputs to the above multidimensional procedures are the "distances" measured by comparing the representations of the different speech signals in the auditory spectral domain.This can be considered as a second generation distortion measure as opposed to the first generation approaches which consider the difference between the acoustic waveforms of the signals as the relevant variable (noise).

The latter part of chapter 2 suggests the way for constructing the third generation measures: The modelling of processes such as pitch and roughness perception indicate that one should be able to measure speech degradation without the need to explicitly use the original signal for the comparison. Instead, relevant parameters from the coded signals (such as roughness)can be monitored and then compared to "standard" values that correspond to an undistorted signal. These values could have been obtained in an initial training phase.

The above procedure is much closer to modelling subjective tests than previous measures: During subjective tests, the human observer can only listen to one coded signal at a time. Therefore one cannot form a difference between this signal and the original signal. One must necessarily follow a procedure similar to the one outlined above in order to assess the degradation.The "training phase" for human observers arises naturally in everyday conversation but can be intensified through the repeated exposure to subjective tests (which establishes the formation of a high dimensionality distortion space). The increased ability of trained subjects in subjective assessment tasks is clearly evident.It is the view of the author that progressive research in the objective assessment of distortion should follow the above line of thought.

A novel application of this new way, to obtain a distortion evaluation,could be in the transmission of speech in a realistic channel environment: Present systems increase the overall bit rate by including bit error detection and correction information for all speech segments. Alternatively, only the corrupted segments that fail a fidelity criterion at the receiver would be retransmitted. A fidelity criterion of this nature could not have knowledge of the original signal and must therefore be of the third generation type as described above. The potential savings in terms of bit rate are obvious.

The transmission scheme just mentioned is an example of variable rate coding. Variable rate coding is investigated in chapter 5. It is shown that significant gains can be easily achieved by exploiting the silent segments in speech. These occur naturally during conversational speech. Additional gains can be achieved,although of a smaller magnitude, by realizing that certain speech sounds are of a noise nature themselves (e.g. frigatives). These types of sound are inherently more tolerant to quantization noise and can therefore be transmitted at a reduced bit rate. Smaller gains still can be made by considering the different tolerance of periodic sounds to rate compression (e.g. in relation to their effective bandwidths). In chapter 5 the procedures to employ a perceptually motivated criterion for variable rate source compression are investigated. The effects of the distribution of distortion are examined, and algorithms for both real-time and voice storage applications are developed. These are illustrated through the use of the multipulse algorithm which provides the capability of varying the bit rate continuously over a wide range of values at medium and low bit rates.

Finally, in chapter 6 a new coding algorithm is developed. The reasons for the development of the algorithm are exemplified below:

At low bit rates it becomes increasingly difficult to ensure an acceptable representation of the speech signal over the entire spectrum. The parts of the signal that suffer most are the interformant valleys and the higher frequency range of the spectrum. This is because speech coders aim to minimize, explicitly or implicitly, the mean square error (mse) between the coded and original signal. In the ideal case, the minimization procedure ensures that the power spectral density (psd) of the error signal is flat across the spectrum. (strictly speaking this is only true where the psd of the signal itself is higher than the psd of the noise). For those frequency regions where this is not true, the psd of the noise follows that of the signal).

From a perceptual viewpoint, the relative powers of signal and noise in each frequency region are also of importance. Therefore,the audibility of noise is higher in the low power regions of the speech signal. (This is related to the "masking" effect of the speech upon coding noise).

Apart from the presence of coding noise in the low power regions of the coded signal, a "muffled" low-pass effect is also present at low bit rates. It is important to realize that this is also the result of minimizing the mse between the coded and original signals: The mse minimization produces a quantization noise which is uncorrelated with the output of the coder. This means that the coded signal power is less than the original signal power, by an amount equal to the coding noise power. This, together with a flat error spectrum, implies that the power of the signal in the high frequency regions is reduced considerably compared to the original signal, which leads to a lowpass effect.

A popular way to compensate for the above effects is to optimize speech coders subject to a weighted mse criterion. This approach produces a mse which is not flat but follows a predetermined shape which, in theory, can be chosen at will to satisfy a predetermined criterion. The total amount of noise cannot of course be reduced. (In fact the overall noise increases).

This method therefore removes noise from the low power frequency regions and places it in the higher power regions of the speech signal spectrum.

Although the above technique has been shown to be beneficial for certain combinations of coder structures and bit rates it has not proved very effective for the latest generation of delayed decision coders (multipulse, CELP etc.) which are designed to operate at very low bit rates. One reason for this is that the noise spectrum cannot be shaped at will, due to the suboptimality of the coders (the mse is only partially minimized, with respect to the model's parameters) and the fact that the speech signal is below the noise level over a considerable part of the spectrum. Another reason is that the lower part of the spectrum is much less tolerant to distortions than the higher part of the spectrum. Thus, although the masking ability of the signal is higher in the lower frequency range, the sensitivity of the perceptual mechanism is also higher: The transfer of noise from the low power, high frequency regions of the spectrum to the high power, low frequency regions is beneficial only in the case of moderate noise levels as in the case of 16 kbits/sec ADPCM.

For low bit rats it is necessary to devise a new approach to speech coding. This approach must ensure a good representation of the perceptually important part of the signal. A good starting point is to follow rate distortion theory and divide the spectrum into a number of passband and stopband regions [3,21, 3-94].

This approach recognizes the fact that for low bit rates it is wasteful to allocate bits to frequency regions that cannot be reproduced with sufficient accuracy. This is taken into account in frequency domain coders [3-21, 3-85] where certain frequency regions are excluded for transmission. In contrast, this fact has not been taken into account in the more recent delayed decision coders [3-70 to 3-74, 3-79, 3-80].

Rate distortion theory in [3-94] deals with stationary signals. The speech signal is a non-stationary source with a time varying frequency spectrum and this must be taken into account when applying a coding technique based on a stopband-passband division of the speech spectrum. In particular it is necessary to conserve:

a) The short-term envelope of the speech signal. The short-term envelope is crucial for phoneme recognition and its distortion leads to "burbling"sounds.

b) The"continuity" in the pitch structure.The pitch structure must somehow be retained even in those regions of the spectrum which are considered as stopbands.

Of course, the passband region(s) must be encoded as efficiently as possible which calls for delayed decision coding to be applied in those regions.

Perceptual distortion, particular to frequency domain coders,arises from their inability to conserve the above features in an efficient manner.

The short-time spectrum can be conserved by transmitting the entire spectral information with the same accuracy irrespective of whether it describes a passband or a stopband region. Efficient coding of the spectral information can be achieved through vector quantization.

With respect to preserving the pitch structure certain psychoacoustic experiments are relevant [2.3-58]

a) When a harmonic signal is shifted up or down in frequency the signal (i) does not sound inharmonic, only a pitch shift is perceived.(ii) The perceived pitch shift is much smaller than the corresponding frequency shift [2.3-60].

b) When two sets of harmonic signals of different fundamentals, one lowpassed and the other high passed are sounded together the pitch of the lowpass signal dominates [2.3-61].

The first steps towards the new algorithm were made in the form of the RELP coder [3-120]. In this coder it is recognised that the long-term speech spectrum has a low pass character. Hence, in accordance with the arguments presented above the speech spectrum is split into a low frequency passband, which is usually a small fraction of the overall spectrum, and a high frequency stopband.

The short term envelope of the whole signal is transmitted in the form of the LPC coefficients. All the remaining bits are allocated to the coding of the passband excitation signal.The pitch structure is preserved even in the upper frequency region at no extra cost in terms of bits. This is achieved by creating images of the baseband across the spectrum through either spectral folding or spectral translation [3-120]. Although this causes the speech harmonics to appear frequency shifted in the spectrum, due to the fact that the baseband is not an integral multiple of the pitch period,this does not lead to any unacceptable degradations as predicted from the above mentioned psychoacoustical experiments.

It is well known that the short term spectrum does not possess the low pass nature of the long term spectrum and that high power formants can be found in the upper frequency region. It is perceptually important to encode the signals within the formant regions with more accuracy than the regeneration procedure allows.This was realized by Un and Lee [3-121] who split the spectrum into several passbands and stopbands. (Two passband regions are used,one from 0 to 500 Hz and the other from 1000-1500 Hz). Although this algorithm provided an improved performance, a more signal specific algorithm can be designed.

The algorithm presented in chapter 6 can be seen as a logical development to the line of thought in [3-120] and [3-121]. A QMF structure is used to split the input signal into 8 contiguous bands.Out of these bands, passbands are selected, whose number and frequency location are made to adapt to the short term speech spectrum of the speech signal. The stopbands are regenerated through a procedure which resembles that of [3-120]. Furthermore, to exploit any remaining correlations across the passbands, these are encoded jointly through a delayed decision analysis by synthesis procedure which is reminiscent of the multipulse algorithm. Thus, the algorithm structure can be considered as a generalized or adaptive RELP. Although the algorithm relies heavily upon rate-distortion theory, its bit compression capability is also based upon a speech compression and speech perception model.

As with any newly-born algorithm, there is considerable scope for improvement in performance: Further work can be done in several parts of the algorithm. Perhaps the most general modification that can be effected is with the coding of the passbands. It is clear that these should be coded in such a way so that their crosscorrelation can be taken into account. Perhaps a stochastic approach can be developed through the VQ procedures outlined at the end of chapter 6. Optimum values can be found for the various algorithm parameters. These parameters that describe the algorithm include the number of filter taps per band for the LPC filters,and the noise shaping factor related to those filters.

Also, the cross correlation between filters from different bands can be exploited. Further improvements can perhaps be realized if the LPC filters are transformed into long delay ("pitch") filters through the addition of an appropriate delay for each filter. These delays are expected to be highly correlated across the bands and therefore require a small number of bits for transmission.

A shortcoming of the present algorithm is the extra delay introduced by the QMF structure. At the moment this delay is added on top of the delay due to the estimation of the filter parameters. One way to avoid the additional delay, introduced by the QMF structure, is to use a system of, one forward, large transform size, DFT followed by many smaller size inverse DFT's to split the signal into pseudo-subband signals as in [3-115]. In this way, the transform delay can be made to overlap completely with the LPC estimation delay.

It is hoped that the work contained in this volume will contribute to further understanding and development in the field of speech quality assessment and coding.

APPENDIX A

Sound Intensity Measurements:

Sound intensity is defined as the power transmitted through a given area in a sound field. In air, for a plane sound wave in a free field, the acoustic intensity is proportional to the square of the amplitude of the pressure deviations caused by the sound wave.

Sound intensity is measured in terms of sound pressure level (SPL) which is the intensity of the sound in dB above the reference level of 10^{-6}Watts/cm^{2} (equivalent to a pressure of 2×10^{-5} N/m^{2}). The reference level was chosen to be close to the human absolute threshold at 1000Hz. Its value is about 6.5 dB SPL.

The following table gives some examples of sound intensity related to speech:

Shouting at close range	100 dB	SPL
Normal conversation	70 dB	SPL
Quiet conversation	50 dB	SPL
Soft whisper	30 dB	SPL

Closely related to SPL is the spectrum level of a sound which is defined as the SPL per unit frequency measured in dBSPL/Hz For noise-like sounds, flat within a frequency range of Δf the two are related by:

$$\text{dB SPL} = \text{dB SPL/Hz} + 10.\log_{10} \Delta f$$

Sometimes, the absolute threshold of a subject for the sound being used is used as a reference. The sound level specified in this way is referred to as sensation level (SL).

The physical intensity corresponding to a given sensation level will differ from subject to subject and from sound to sound.

More information can be found in the International Standards Organization publication ISO R131

APPENDIX B

Octave Band Filters and Derivatives

The spectral density of a sound is the power in a 1 Hz band. Usually an octave filter, one-third octave or occasionally a one-tenth octave filter is used for measurements. The center frequencies are conventionally set to certain values. These are shown in table B1.

There are three one-third octave bands within an octave band. The centre frequency f_c is half way between the upper fu and lower f_1 cutoffs on a logarithmic frequency scale:

$$\log f_c = ½ \log f_1 + ½ \log f_u$$

or

$$f_c^2 = f_1 \; f_u$$

for an octave band: $f_u = 2f_1$, $f_u - f_1 = f_1$

$$f_c^2 = f_u f_1 = 2f_1^2$$

$$f_c = \sqrt{2} f_1 = f_u / \sqrt{2}$$

for a 1/3 octave band: $f_u = \sqrt[3]{2} f_1$, $f_c = \sqrt{\sqrt[3]{2}} \; f_1$

and the ratio between successive centre frequencies of contiguous filters is given by

$$\frac{fc_2}{fc_1} = \sqrt{\frac{2}{\sqrt[3]{2}}} = \sqrt[3]{2}$$

Table B–T1 [a]
Preferred Center Frequencies of Octave and One-Third Octave Filter

One-third octave	Octave
	16
20	
25	
	31.5
40	
50	
	63
80	
100	
	125
160	
200	
	250
315	
400	
	500
630	
800	
	1,000
1,250	
1,600	
	2,000
2,500	
3,150	
	4,000
5,000	
6,300	
	8,000
10,000	
12,500	
	16,000

APPENDIX C

C1 Positive Integer Constraints

The solution of 3.2-162 can result in negative values for R_K when the overall bit rate R is sufficiently low. The bands with negative bits are not transmitted and are therefore given 0 bits. The bit rate of the remaining bands has to be readjusted so that the total equals NR (equation 3.2-158). Consider equation 3.2-162 again. Let a and b be two bands of sufficiently high σ_{xK}^2 to receive a positive number of bits, then

$$R_a = R + \frac{1}{2} \log_2 \frac{\sigma_a^2}{p} \qquad C.1$$

where

$$p = \left[\prod_{l=1}^{N} \sigma_{xl}^2 \right]^{1/N} \qquad C.2$$

and

$$R_b = R + \frac{1}{2} \log_2 \frac{\sigma_b^2}{p} \qquad C.3$$

from C.1 and C.3 $\quad R_a - R_b = \frac{1}{2} \log_2 \frac{\sigma_a^2}{\sigma_b^2} \qquad$ C.4

Assume now that a-priori information exists about which bands will receive a positive number of bits.

The aim is to distribute the available bits among these bands only. Let bands l=1,2....M be those bands with sufficiently high σ_{xK}^2 to receive positive bits. The optimum bit allocation formula is again that of C.1 and C.3, with C.2 replaced by

$$P' = [\prod_{l=1}^{M} \sigma_{xl}^{2}]^{1/M} \qquad C.5$$

therefore

$$R_a = R + \frac{1}{2} \log_2 \frac{\sigma_a^2}{P'} \qquad C.6$$

$$R_b = R + \frac{1}{2} \log_2 \frac{\sigma_b^2}{P'} \qquad C.7$$

from which it follows that

$$R_a - R_b = \frac{1}{2} \log_2 \frac{\sigma_a^2}{\sigma_b^2} \qquad C.8$$

It can be seen that equtions C.4 and C.8 are identical, and desribe an optimum bit allocation for bands a and b. This leads to an iterative algorithm for the new bit alocation:

1. Use equation 3.2-162 to find the R_K's. This step establishes the relationship given by equations C.1 and C.3

2. Sum the number of negative allocated bits. Set the rates R_K of these bands with negative allocations to zero. Divide the number of negative bits equaly amongst the remaining bands. The equal division of bits guarantees that equations C.4 (or C.8) are not violated hence an optimum bit allocation is in effect.

3. If any of the remaining bands with non zero bit allocation have now a negative bit allocation go to step 2 else exit.

Note that since the above allocation is stil optimum, the noise power wil be flat across those bands that do eventually receive a positive number of bits.

Usually it is also necessary to restrict the R_k to integer values. One procedure to obtain an integer allocation close to the optimum allocation is as follows:

1. Find the R_k which is closest (but not equal) to an integer value. Round it up to that integer value and divide the difference between the new and old rate equally amongst those bands which do not haVe an integer allocation.

2. If all R_k are integers exit else go to 1.

One simplified procedure due to Ramstad [reference 95 of chapter 3] stems again from equation C.4: If one sets

$$R_a - R_b = 1 \qquad \text{C.9}$$

then

$$\frac{1}{2}\log_2 \frac{\sigma_a^{\ 2}}{\sigma_b^{\ 2}} = 1 \qquad \text{C.10}$$

or

$$\sigma_a/\sigma_b = 2 \qquad \text{C.11}$$

Therefore 1 bit is worth a halving of the corresponding σ_k.

Ramstad's method is as follows:

1. Start from the maximum σ_K. Allocate 1 bit to this band and halve σ_K.
2. Subtract one bit from the total bit rate.
3. If all bits are exhausted exit, else go to 1

The two different procedures mentioned above were run for a large number of blocks. Identical results were obtained.

Noise weighting

It was shown that the optimum bit allocation leads to a flat noise spectrum. An optimum bit allocation subject to a weighted m.s.e. criterion is obtained if the σ_{xK}^2 are replaced by $W_K\sigma_{xK}^2$. The flat σ_{qK}^2 rule is then replaced by a flat $W_K\sigma_{qK}^2$ rule. A graphical comparison of weighted and unweighted noise is shown in figure C.1.

A proposed class of weighting functions is [ref 21 of chapter 3]

$$W_k = (\sigma_{xk}^2)^{\lambda} \qquad \text{C.12}$$

λ = o gives a white error spectrum, ($\sigma_{qK}^2 = \sigma_{ql}^2$ any k,l) whereas $\lambda = -1$ gives a constant bit assignment ($R_K = R_l$ any k,l). In the latter case the error spectrum has the shape of the input signal. The noise spectrum for various values of λ is shown schematically in figure C.2.

With the weighting function of equation C.12, equation C.4 is modified to

$$R_a - R_b = \frac{1}{2}\log_2 \frac{W_a\sigma_a^2}{W_b\sigma_b^2} \qquad \text{C.13}$$

or

$$R_a - R_b = \frac{1}{2} \log_2 \left(\frac{\sigma_a^2}{\sigma_b^2}\right)^{\lambda+1} \qquad \text{C.14}$$

and for

$$R_a - R_b = 1 \qquad \text{C.15}$$

$$\log_2 \left(\frac{\sigma_a}{\sigma_b}\right)^{\lambda+1} = 1 \qquad \text{c.16}$$

or

$$\sigma_a/\sigma_b = 2^{\frac{1}{\lambda+1}} = r \qquad \text{C.17}$$

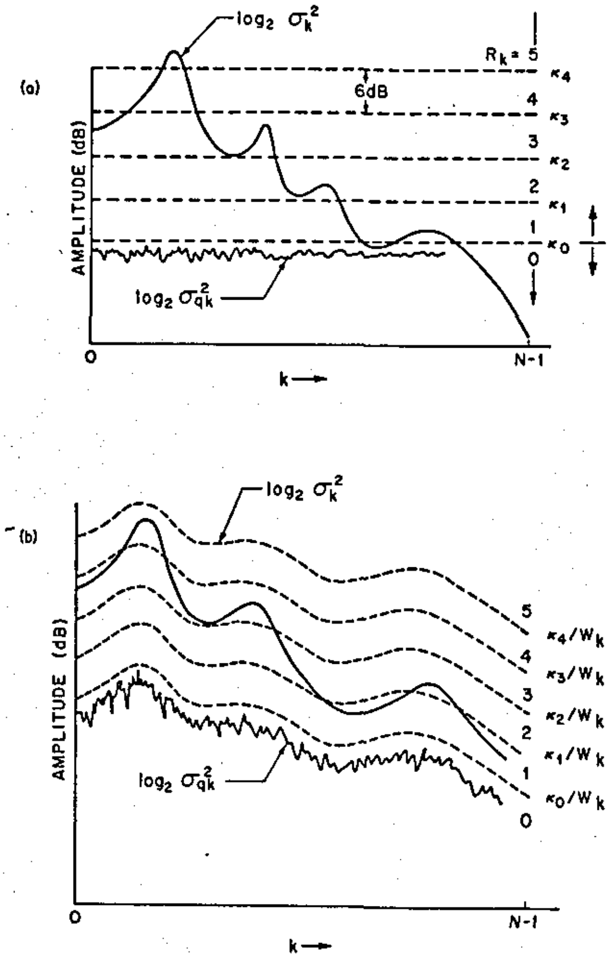
r is now Ramstad's ratio and equation

$$2^{\frac{1}{\lambda+1}} = r \qquad \text{C.18a}$$

or

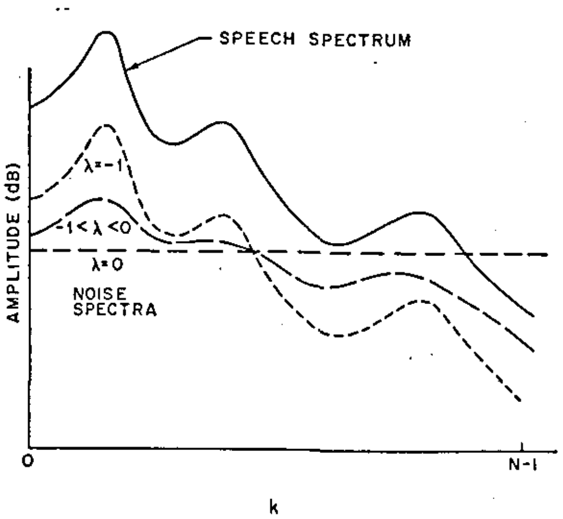
$$\lambda = -\log_2 r - 1 \qquad \text{C.18b}$$

gives the equivalence formula for the two types of weighting.



A graphical illustration of bit allocation for obtaining (a) white and (b) frequency-weighted quantization noise spectra

Figure C1 [Ref. 21, Ch 3.2]



The psd of transform coding noise as a function of error-weighting parameter λ

Figure C2 [Ref. 21, Ch 3.2]

APPENDIX D

Solution to the following (constrained) minimization problem:

Minimize

$$\sigma_q^2 = \frac{1}{N} \sum_{K=o}^{N-1} \sigma_{qk}^2 \qquad \text{D.1}$$

subject to

$$R = \frac{1}{N} \sum_{k=o}^{N-1} R_k = \text{constant} \qquad \text{D.2}$$

Using lagrange multipliers the solution can be found by setting

$$\frac{\partial}{\partial R_k} [\sigma_q^2 - \lambda (R - \frac{1}{N} \sum_{k=o}^{N-1} R_k)] = o$$

for k=0,1,.....N-1 $\qquad$ D.3

The conventional formula for σ_{qk}^2 is given by

$$\sigma_{qk}^2 = \epsilon_*^2 2^{-2R_k} \sigma_{xk}^2 \qquad \text{D.4a}$$

which can be written as

$$\frac{\sigma_{xk}^2}{\sigma_{qk}^2} = \frac{1}{\epsilon_*^2} 2^{2R_k} \qquad \text{D.4b}$$

or

$$10\log_{10}\frac{\sigma_{xk}^{2}}{\sigma_{qk}^{2}} = 10\log_{10}\frac{1}{\epsilon_{*}^{2}} + 20R_{k}.\log_{10}2 \qquad \text{D.4c}$$

giving

$$\mathrm{SNR(dB)} = A + 6R_{k} \qquad \text{D.4d}$$

which is usually termed the 6dB/bit rule.

Under more general assumptions the constants in equation D4d can be taken to be functions of the subband index:

$$\mathrm{SNR(dB)} = A_{k} + a_{k}R_{k}(20\log_{10}2) \qquad \text{D.5}$$

from which the simplified form D4d is obtained by setting

$$A_{k} = A = \text{constant} \qquad \text{D6a}$$

and

$$a_{k} = 1 \qquad \text{D6b}$$

D5 can also be written as

$$10\log_{10}\frac{\sigma_{xk}^{2}}{\sigma_{qk}^{2}} = 10\log_{10}\frac{1}{\epsilon_{*k}^{2}} +2\ R_{k}a_{k}[10\log_{10}2] \qquad \text{D.7}$$

where

$$\frac{-A_{k}}{10} = \log_{10}\epsilon_{*k}^{2} \qquad \text{D8a}$$

or

$$10^{\frac{-A_k}{10}} = \epsilon_{*k}^{2} \tag{D8b}$$

From D7

$$\frac{\sigma_{xk}^{2}}{\sigma_{qk}^{2}} = \frac{1}{\epsilon_{*k}^{2}} \, 2^{2a_k R_k} \tag{D9a}$$

and

$$\sigma_{qk}^{2} = \epsilon_{*k}^{2} \, 2^{-2a_k R_k} \, \sigma_{xk}^{2} \tag{D9b}$$

The purpose of this appendix is to provide the solution to the minimization problem given by conditions D1 and D2 provided that σ_{qk}^2 is given by D9b.

Substituting D9b into D3 gives
(all summations are carried out between k=o and k=N-1)

$$\frac{\partial}{\partial R_k} \left[\frac{1}{N} \Sigma \, \epsilon_{*k}^{2} 2^{-2a_k R_k} \sigma_{xk}^{2} - \lambda \left(R - \frac{1}{N} \Sigma R_k\right) \right] = 0 \tag{D10}$$

giving

$$\frac{1}{N} \epsilon_{*k}^{2} \, 2^{-2a_k R_k} (\ln 2)(-2a_k) \sigma_{xk}^{2} + \frac{\lambda}{N} = 0 \tag{D11}$$

therefore

$$\lambda = 2a_k \epsilon_{*k}^{2} \, (\ln 2) \, \sigma_{xk}^{2} \, 2^{-2a_k R_k} \tag{D12}$$

let

$$B_k = 2a_k\epsilon_{*k}^{2}(\ln 2) \tag{D13}$$

then from D12

$$\log_2\lambda = \log_2 B_k + \log_2\sigma_{xk}^{2} - 2a_kR_k \tag{D14}$$

and

$$\frac{1}{a_k}\log_2\lambda = \frac{1}{a_k}\log_2 B_k + \frac{1}{a_k}\log_2\sigma_{xk}^{2} - 2R_k \tag{D15}$$

taking an average over the subbands and using D2

$$\langle\log_2\lambda\rangle\frac{1}{N}\Sigma\frac{1}{a_k} = \frac{1}{N}\Sigma\frac{\log_2 B_k}{a_k} + \frac{1}{N}\Sigma\frac{\log_2\sigma_{xk}^{2}}{a_k} - 2R \tag{D16}$$

which gives

$$\log_2\lambda = \frac{1}{\frac{1}{N}\Sigma\frac{1}{a_k}}\left[\frac{1}{N}\Sigma\frac{\log_2 B_k}{a_k} + \frac{1}{N}\Sigma\frac{\log_2\sigma_{xk}^{2}}{a_k} - 2R\right] \tag{D17}$$

The left hand sides of D14 and D17 are equal so the r.h.s. can be equated:

$$\log_2 B_k + \log_2\sigma_{xk}^{2} - 2a_kR_k =$$

$$\frac{1}{\frac{1}{N}\Sigma\frac{1}{a_k}}\left[\frac{1}{N}\Sigma\frac{\log_2 B_k}{a_k} + \frac{1}{N}\Sigma\frac{\log_2\sigma_{xk}^{2}}{a_k} - 2R\right] \tag{D18}$$

and dividing by $-2a_k$,

$$\frac{-\log_2 B_k}{2a_k} - \frac{\log_2 \sigma_{xk}^2}{2a_k} + R_k =$$

$$\frac{1}{\frac{-2a_k}{N}\Sigma\frac{1}{a_k}}\left[\frac{1}{N}\Sigma\frac{\log_2 B_k}{a_k} + \frac{1}{N}\Sigma\frac{\log_2 \sigma_{xk}^2}{a_k} - 2R\right] \qquad D19$$

therefore the optimum bit rate for band K, R_k is given by

$$R_k = \frac{-1}{2a_k}\left[\frac{1}{\frac{1}{N}\Sigma\frac{1}{a_k}}\left(\frac{1}{N}\Sigma\frac{\log_2 B_k}{a_k} - 2R + \frac{1}{N}\Sigma\frac{\log_2 \sigma_{xk}^2}{a_k}\right)\right.$$

$$\left. -\log_2 B_k - \log_2 \sigma_{xk}^2\right] \qquad D20$$

setting $B_K=B$ and $a_K=1$ for all k the conventional formula is derived from D20:

$$R_k = R + \frac{1}{2}\log_2 \frac{\sigma_{xk}^2}{\left[\prod_{k=o}^{N-1}\sigma_{xk}^2\right]^{1/N}} \qquad D21$$

In practice, additional constraints to the one given by D2 have to be imposed upon the solution in D20: first the solution R_K must be positive since a negative bit allocation is meaningless. Second, R_K is usually required to take ony integer values.

Two algorithms have been deviced to deal with these two additional constraints. The procedure for dealing with negative bit (R_K) allocation is as follows:

1) For any $R_k<0$ set $R_k=o$

Step 1 has increased the average bit rate to a new value R' The increase in bits ΔR is given by ΔR = N(R'-R). Step 2 brings the average bit rate back to the required value of R.

2) Remove a certain amount of bits from each band for which R_k>o so that the new average bit rate is euqal to R

3) If any of the R_k's are less than zero go to 1 else EXIT.

Procedure for integer bit allocation:

1) Round that R_k nearest to an integer value to that integer value (but skip if zero)

As in the previous procedure step one alters the average bit rate. Step 2 corrects for step 1.

2) Distribute the excess bit rate ΔR (which may be positive or negative) amongst the non-integer R_k according to some optimum way.

3) If all R_k's are integers exit otherwise go to 1

It is clear that step 2 in both procedures requires a distribution of an excess bit rate ΔR (which may be positive or negative) in an optimal way amongst a subset of the bands. To find the new optimum bit allocation subject to the additional constraints consider the following:

The weighted (by $2a_k$) difference in bitrates between any two bands (1,2) is given by (from D20)

$$2a_1R_1-(2a_2R_2)=\log_2B_1+\log_2\sigma_{x1}^{2}-(\log_2B_2+\log_2\sigma_{x2}^{2}) \qquad \text{D22}$$

or

$$a_2R_2-a_1R_1 = \frac{1}{2} [\log_2 \frac{\sigma_2^{\;2}}{\sigma_1^{\;2}} + \log_2 \frac{B_2}{B_1}] \qquad \text{D23}$$

let the r.h.s. of D23 be denoted by A_{12}. A_{12}, the weighted difference, does not depend on any characteristics of any bands other than (1,2) or the average overall bit rate. This point is elaborated in more depth in appendix C. Assume that a new amount of bits ΔR_K has to be added to each band, bringing the individual bit rates to new values R'_K. For an optimum distribution of bits

$$a_2R'_2 - a_1R'_1 = A_{12} \qquad \text{D24}$$

or

$$a_2R_2 - a_1R_1 + a_2\Delta R_2 - a_1\Delta R_1 = A_{12} \qquad \text{D25}$$

which implies (from D23) that

$$a_2\Delta R_2 - a_1\Delta R_1 = 0 \qquad \text{D26a}$$

or

$$a_2\Delta R_2 = a_1\Delta R_1 = C \qquad \text{D27a}$$

and for any band K

$$\Delta R_k = \frac{C}{a_k} \qquad \text{D27c}$$

let ΔR be the total excess bit rate

$$\Delta R = \sum_L \Delta R_k = C \sum_L \frac{1}{a_k} \qquad \text{D28}$$

where the summation is over the subset L of those bands eligible for redistribution according to the two reallocation procedures set above.

Hence

$$C = \frac{\Delta R}{\sum_L \frac{1}{a_k}} \qquad \text{D29}$$

and, finally

$$\Delta R_k = \frac{\Delta R}{a_k \sum_L \frac{1}{a_k}} \qquad \text{D30}$$

which gives the amount by which each bit rate R_k has to be modified by in step 2 in the two reallocation procedures.

The solution to equations D20 and D30 requires knowledge of the constants a_k and B_k, or equivalently a_k and ϵ_{*k}^2. These can be obtained by fitting straight lines to the function $10\log_{10}[\sigma_{xk}^2/\sigma_{qk}^2]$ by evaluating the l.h.s. of D7 per block and then averaging these values over time i.e.

$$\frac{1}{M}\sum_M 10\log_{10}\frac{\sigma_{xk}^2}{\sigma_{qk}^2} = \frac{1}{M}\sum_M 10\log_{10}\frac{1}{\epsilon_{*k}^2} + \frac{20\log_{10}2R_k}{M}\sum_M a_k$$

where M is the total number of time blocks D31

An estimate for a_k was obtained from

$$\hat{a}_k = \frac{1}{M}\sum_M a_k \qquad \text{D32}$$

and an estimate for ϵ_{*k}^{2} from

$$10\log_{10} \frac{1}{\hat{\epsilon}_{*k}^{2}} = \frac{1}{M} \sum_{M} 10\log_{10} \epsilon_{*k}^{2} \qquad \text{D33a}$$

or

$$\hat{\epsilon}_{*k}^{2} = [\prod_{M} \epsilon_{*k}^{2}]^{\frac{1}{M}} \qquad \text{D33b}$$

The experimental data corresponding to D31 do not necessarily describe a straight line. Hence, least squares straight-line fitting procedures were actualy used to obtain a_k andϵ_{*k}^{2}.

APPENDIX E

Proof to equation 4.6:

From equation 4.1b

$$2^{2R} = \frac{\sigma_x^{2}}{D_w} \tag{E.1}$$

$$D_w = 2^{-2R}\sigma_x^{2} = 2^{-2R}\,\frac{1}{F}\int_o^F S_x(f)df \tag{E.2}$$

$$10\log_{10}D_w = 10\log_{10}2^{-2R} + 10\log_{10}[\frac{1}{F}\int_o^F S_x(f)df] \tag{E.3}$$

from equation 4.5:

$$R = \frac{1}{F}\int_o^F \frac{1}{2}\log_2[\frac{S_x(f)}{D}]\,df \tag{E.4}$$

$$\therefore\ -2R = \frac{1}{F}\int_o^F -\log_2 S_x(f)df + \log_2 D \tag{E.5}$$

and

$$2^{-2R} = 2^{[\frac{1}{F}\int_o^F -\log_2 S_x(f)df \,+\, \log_2 D]} \tag{E.6}$$

$$\therefore\ 10\log_{10}2^{-2R} = [\frac{1}{F}\int_o^F -\log_2 S_x(f)df+\log_2 D].\log_{10}2 \tag{E.7}$$

and since

$$\log_2 a = \frac{\log_{10} a}{\log_{10} 2} \qquad \text{E.8}$$

E.7 can be written as:

$$10\log_{10} 2^{-2R} = -\frac{1}{F} \int_{o}^{F} 10\log_{10} S(f)df + 10\log_{10} D \qquad \text{E.9}$$

substituting $10\log_{10} 2^{-2R}$ from E.9 into E.3, equation 4.6 is obtained.

APPENDIX F

The Multipulse algorithm

In the multipulse algorithm the excitation is defined in terms of pulse amplitudes and locations.

The excitation parameters are obtained through an explicit (weighted) error minimization procedure as shown in figure F1a. The error, formed by subtracting the coded from the original signal is weighted by a filter W(Z) given by

$$W(Z) = \frac{A(Z)}{A(\frac{Z}{g})} \qquad \text{F.1}$$

where A(Z) is the appropriate inverse filter derived from a block of speech which ideally contains the multipulse minimization frame. The parameter g controls the degree of noise shaping. Appropriate values are around 0.8-0.9. Figure F1b is functionally equivalent to figure F1a since it is obtained from F1a through permissible linear operations. In figure F1b the weighting filter has been removed from the minimization loop which leads to faster implementation compared to the structure in figure F1a.

The minimization is performed in a block mode: Without any error weighting (i.e. g = 1), the error signal e(n) in one block containing the samples n=1,2,....N is given by

$$e(n) = X'(n) - [\sum_{i=1}^{I} g_i h(n-m_i) + M_n] \qquad \text{F.2}$$

X'(n) is the input signal whereas the terms in the brackets constitute the synthesized signal, when a total of I pulses are used for modelling the excitation. The parameter g_i is the ith pulse amplitude at location m_i,

$h(n)$ is the impulse response of $1/A(Z)$ and M_n the contribution of the excitation from the previous frame to the current synthesized speech block.

Let

$$X(n) = X'(n) - M(n) \qquad \text{F.3}$$

Since both $X'(n)$ and $M(n)$ are independent of m_i and g_i, they can be lumped together into one term. $M(n)$ is the output of the synthesis filter $1/A(Z)$ with zero input, provided the last synthesized values from the previous block are used to form the filter's memory.

The weighted noise signal $e_w(n)$ can be obtained by passing $e(n)$ through the weighting filter $W(Z)$ given in F1. This is of course equivalent to passing every component of the r.h.s. of equation F.2 through the weighting filter (distributive property of convolution), which is schematically shown in figure F1b. Therefore the weighted noise is given by

$$e_w(n) = X_w(n) - \sum_{i=1}^{I} g_i h_w(n-m_i) \qquad \text{F.4}$$

$$n=1,2....N$$

where h_w is the impulse of $1/A(Z/g)$.

The error minimization procedure attempts to minimize the short term noise power P_I given by

$$P_I = \sum_{n=1}^{N} e_w^{\,2}(n) \qquad \text{F.5a}$$

where the subscript I indicates that P is the error energy resulting from a model employing I pulses. Hence

$$P_I = \sum_{n=1}^{N} [X_w(n) - \sum_{k=1}^{I} g_i h_w(n-m_i)]^2 \qquad \text{F.5b}$$

Assuming that the locations of the pulses are already known the error power can be minimized w.r.t. the pulse amplitudes:

$$\frac{\partial P_I}{\partial g_k} = \sum_{n=1}^{N} \frac{\partial}{\partial g_i} [X_w(n) - \sum_{i=1}^{I} g_i h(n-m_i)]^2$$

$$=2 \sum_{n=1}^{N} [X_w(n) - \sum_{i=1}^{I} g_i h_w(n-m_i)] h(n-m_k) = 0$$

$$k=1,2.....I \qquad \text{F.6}$$

$$\therefore \sum_{n=1}^{N} X_w(n) h_w(n-m_k) = \sum_{n=1}^{N} \sum_{i=1}^{I} h(n-m_k) g_i h_w(n-m_i)$$

$$k=1,2,......I \qquad \text{F.7}$$

Expanding P_I in F.5:

$$P_I = \sum_{n=1}^{N} X_w^2(n) -2 \sum_{n=1}^{N} X_w(n) \sum_{i=1}^{I} g_i h_w(n-m_i)$$

$$+ \sum_{n=1}^{N} [\sum_{i=1}^{I} g_i h_w(n-m_i)]^2 \qquad \text{F.8}$$

From F.7, multiplying by g_k and summing from k=1 to k=I

$$\sum_{i=1}^{I} g_k \sum_{n=1}^{N} X_w(n) h_w(n-m_k) = \sum_{n=1}^{N} \sum_{k=1}^{I} g_k h(n-m_k) g_i h_w(n-m_i) \qquad \text{F.9}$$

or

$$\sum_{n=1}^{N} X_w(n) \sum_{k=1}^{I} g_k h_w(n-m_k) = \sum_{n=1}^{N} [\sum_{i=1}^{I} g_i h_w(n-m_i)]^2 \qquad \text{F.10}$$

the term on the l.h.s. of equation F10 is half the middle term of F.7 whilst the r.h.s. of equation F.10 is the last term of F.7. Substituting F.10 into F.7 obtains:

$$P_I = \sum_{n=1}^{N} X_w^2(n) - \sum_{n=1}^{N} [\sum_{i=1}^{I} g_i h_w(n-m_i)]^2 \qquad F.11$$

Note that equation F.4 can be considered to describe the error between a modified speech signal $X_\omega(n)$ and its synthesized counterpart. In this respect P_I is equal to the power of the original modified signal minus the power of the synthesized (modified) signal.

It follows from F.11 that the power of the synthesized (modified) signal is always less than the power of the original (modified) signal by an amount equal to the error power P_I.

A solution of the system of equations described by F.7 gives the optimum pulse amplitudes once the positions are known. Alternatively P_I in F.11 can be evaluated through F.7 for every possible set of locations $\{m_i, i = 1,2....I\}$. The best set of locations that fully minimizes P_I can then be chosen. This would result in an exhaustive search which is not practical in terms of computational cost.

An iterative approach can be derived by writing equation F.4 as

$$e_w^I(n) = X_w(n) - \sum_{i=1}^{I-1} g_i h_w(n-m_i) - g_I h_w(n-m_I) \qquad F.12$$

The superscript denotes that $e_\omega^I(n)$ is the error resulting from modelling with I pulses. The first two terms of the r.h.s. of F.12 represent the error resulting from modelling with I-1 pulses, hence

$$e_w^{I}(n) = e_w^{I-1}(n) - g_I h_w(n-m_I) \qquad \text{F.13}$$

F.5 can now be written as

$$P_I = \sum_{n=1}^{N} [\, e_w^{I-1}(n) - g_I h_w(n-m_I)]^2 \qquad \text{F.14}$$

P_I can now be minimized w.r.t. g_I which is the solution to

$$\frac{\partial P_I}{\partial g_I} = \sum_{n=1}^{N} [\, e_w^{I-1}(n) - g_I h_w(n-m_I)] h_w(n-m_I) = 0 \qquad \text{F.15}$$

hence

$$g_I = \frac{\sum_{n=1}^{N} e_w^{I-1}(n) h_w(n-m_I)}{\sum_{n=1}^{N} [h_w(n-m_I)]^2} \qquad \text{F.16}$$

Therefore for a given location m_I the optimum amplitude, g_I (assuming that all $\{g_i,\ i=1,2....I-1\}$ have already been determined) is given by F.16.

Expanding F.14 gives

$$P_I = \sum_{n=1}^{N} [e_w^{I-1}(n)]^2 - 2g_I \sum_{n=1}^{N} [e_w^{I-1}(n) h_w(n-m_I)]$$

$$+ g_I^2 \sum_{n=1}^{N} [h_w(n-m_I)]^2 \qquad \text{F.17}$$

using F.16 to substitute for the multiplier of $-2g_I$ in F.17 obtains:

$$P_I = \sum_{n=1}^{N} [e_w^{I-1}(n)]^2 - g_I^2 \sum_{n=1}^{N} [h_w(n-m_I)]^2 \qquad \text{F.18}$$

or, following the subscript notation for power

$$P_I = P_{I-1} - g_I^2 \sum_{n=1}^{N} [h_w(n-m_I)]^2 \qquad \text{F.19}$$

In F.19 the value for P_{I-1} and, in the case of the autocorrelation approximations, (equation F26) the multiplier of g_I^2, are independent of the location m_I. Therefore the minimum value for P_I is obtained when g_I^2 (or $|g_I|$) is a maximum. (In the case of the autocovariance solution, the multiplier of g_I^2 in F19 has to be included into the term that must be maximized).

A way to determine the best amplitude and position for pulse I is therefore to calculate F16 for each and every location m_I and chose that m_I which maximizes $|g_I|$. After this pulse is defined one can then proceed to pulse I+1 and so on. Of course

$$e_w^0(n) = X_w(n)$$

$$P_0 = \sum_{n=1}^{N} X_w^2(n) \qquad \text{F.20}$$

At every stage of the algorithm equations F.7 can be applied to jointly optimize the amplitudes for the already known positions, improving the quality at the expense of more complexity.

From F.19 expanding every P_i

$$P_I = \sum_{n=1}^{N} X_w^2(n) - \sum_{i=1}^{I} g_i^2 \sum_{n=1}^{N} [h_w(n-m_i)]^2 \qquad \text{F.21a}$$

or

$$P_I = \sum_{n=1}^{N} X_w^2(n) - \sum_{n=1}^{N} \sum_{i-1}^{I} [g_i h_w(n-m_i)]^2 \qquad \text{F.21b}$$

which is in a similar form as F.11. Note that the $\{g_i\}$ are different.

To find the maximum g_I the function

$$A^{I}(m_i) = \sum_{n=1}^{N} e_w^{I-1}(n)h_w(n-m_i) \qquad F.22$$

must be calculated for each $m_i=1,2....N$. Note that the denominator of F.16 need only be calculated once for each of $\{m_i, i=1,2....N\}$. (Note that this term is the same as the multiplier of g_I^2 in F19).

Equation F.22 can be written as

$$A^{I}(m_i) = \sum_{n=1}^{N} [e_w^{I-2}(n)-g_{I-1}h_w(n-m_{I-1})]h_w(n-m_i)$$

$$= \sum_{n=1}^{N} e_w^{I-2}(n)h_w(n-m_i)-g_{I-1} \sum_{n=1}^{N} h_w(n-m_{I-1})h_w(n-m_i)$$

$$=A^{I-1}(m_i)-g_{I-1} \sum_{n=1}^{N} h_w(n-m_{I-1})h_w(n-m_i) \qquad F.23$$

$$\text{With } A^{1}(m_i) = \sum_{n=1}^{N} X_w(n)h_w(n-m_i) \qquad F.24$$

Expanding $A^{I-1}(m_i)$, $A^{I-2}(m_i)...A^{2}(m_i)$ in a similar way in F.23 then obtains

$$A^{I}(m_i)= \sum_{n=1}^{N} X_w(n)h_w(n-m_i) - \sum_{i=1}^{I-1} g_k \sum_{n=1}^{N} h_w(n-m_k)h_w(n-m_i) \qquad F.25$$

Therefore the numerator of F.16 can be evaluated iteratively without the need to evaluate $e_\omega(n)$ explicitly.

The algorithm can be simplified considerably if autocovariance estimates are replaced by autocorrelation estimates:

$$Rh_wh_w(|m_k-m_i|) = \sum_{n=0}^{N-|m_i-m_k|-1} h_w(n)h_w(n+|m_k-m_i|) \qquad F.26$$

to replace the multiplier of $\sum_{k=1}^{I-1} g_k$ in F.25. Also

$$Rh_wX_w(m_i) = \sum_{n=m_i}^{N} X_w(n)h_w(n-m_i) \qquad F.27a$$

or equivalently

$$Rh_wX_w(m_i) = \sum_{j=o}^{N-m_i} X(j+m_i)h_w(j) \qquad F.27b$$

(by letting $j = n - m_i$) to replace the first term of the r.h.s. of F.25

Autocorrelation estimates can also be used in F.7, F16 and F19.

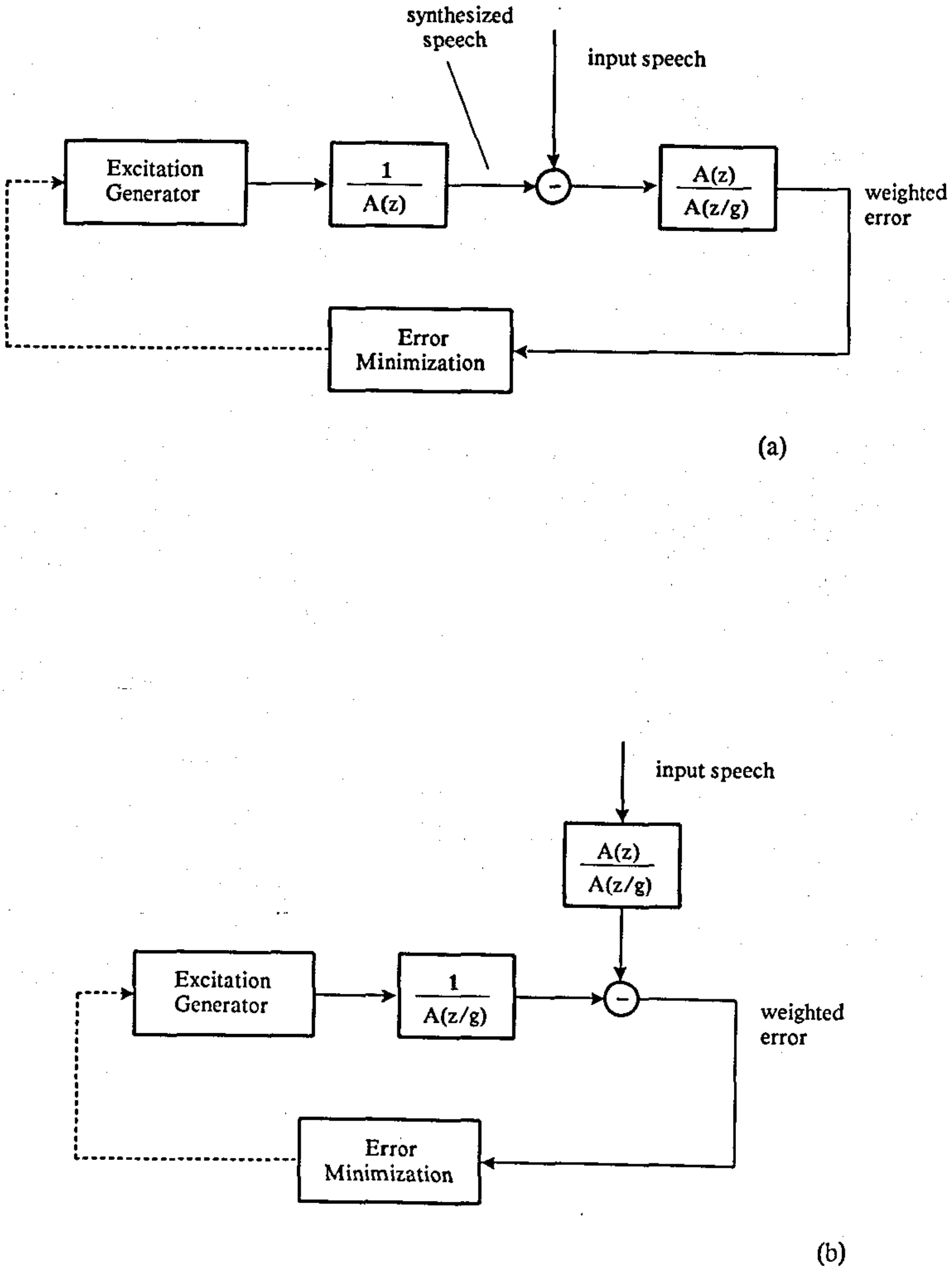


Figure F1 The multipulse algorithm.
——— Linear part.
--------- Nonlinear part.

Let $R = \sum_{i} \Delta R_i$

R is the total number of bits required to increase the number of pulses in each band by one.

The aim is to minimize the sum of E and R i.e.

minimize $E + R = \sum_{i} (\Delta E_i + \Delta R_i)$

This will ensure that the pulse allocation results in a bit allocation as near to the optimum as possible.

A two stage iterative algorithm was used for the minimization. This is given below:

```
START1: From those bands for which $\Delta E_i \geq \Delta R_i$,
        (if none go to START2) find that band m
        with the minimum $\{\Delta R_i\} = \Delta R_m$

        Now,
        $\Delta E_m + \Delta R_m = FSX_m - PX_m + X_m - FSX_m$

                        $= X_m - PX_m$

        This is the number of bits required to increase
        the pulse allocation from $P_m$ to $P_m + 1$

        $\Delta E_m + \Delta R_m$ = Bits $(P_m + 1)$ - Bits $(P_m)$

        IF $(E \geq \Delta E_m + \Delta R_m)$ THEN
             set the value of $P_m$ to $P_m + 1$
             set the value of E to $E - (\Delta E_m + \Delta R_m)$
             set the value of $\Delta E_m$ to $-\Delta R_m$

             band m is now excluded

             GO TO START1

        END IF
```

```
START2: Find the band K with the maximum $\{\Delta E_i\}=\Delta E_k$
        IF $(E>Bits(P_k+1)-Bits(P_k))$ THEN
            set the value of $P_k$ to $P_k+1$
            set the value of E to $E-[Bits(P_k+1) -Bits (P_k)]$
            set the value of $\Delta E_k$ to $\Delta E_k-[Bits(P_k+1) -Bits (P_k)]$
        END IF
        GO TO START2: (Continue for a number of interations e.g. 20)
```

APPENDIX H: A COMPARATIVE STUDY OF CODER COMPLEXITIES

The following table provides a complexity comparison between the simplified multipulse algorithm based on the autocorrelation approximation (reference 80 of chapter 3) and the final algorithm of chapter 6.

A predictor order P = 12 is assumed for the multipulse algorithm whereas P = 4 for the new algorithm. A multipulse frame of N = 128 samples is assumed. This corresponds to an equivalent frame of N/8 for the subband algorithm. The LPC analysis frame is taken to be twice a long as the multipulse frame in both cases, namely 2N and N/4. The values are given in terms of multiply/add operations per multipulse frame. These are expressed as a function of N and therefore the multipliers of N give the number of multiply/add operations per sample, where the sampling rate is 8KHz.

The following abbreviations are related to specific subroutines common to both schemes:

Burg: The homonymous lattice formulation procedure for calculating the predictor coefficients.

IMRESP: The procedure for evaluating the fist N (N/8) impulse response terms of the LPC filter.

AUTO: The procedure for evaluating the first N (N/8) autocorrelation terms of the impulse response of the LPC filter.

SUBMEM: Subtration of the filter's memory from the current frame

WEIGHT: ARMA weighting by the filter [1-P(Z)]/[1-P(Z/a)] where P(Z) is the LPC predictor and a the noise weighting factor.

CROSS: To calculate the first N terms of a crosscorrelation function (between the impulse response of the LPC filter and the modified speech signal)

Routine	General Formula	Multipulse	New coder
Burg	5x2NxP	5x2Nx12 =120N	5x32x4 =4480=35N
calculated every two frames (cetf):	5xNxP	60N	17.5N
IMRESP	PxN		
cetf:	Px(N/2)	12x(128/2) =6xN	4x(16/2)x7 =1.75N
AUTO (cetf)	Px(N/2)	6xN	1.75N
SUBMEM	PxN	12xN	3.5N
WEIGHT	2xPxN	24xN	7xN
CROSS	PxN	12xN	3.5xN

PROCEDURE	MULTIPULSE	NEW CODER
During the pulse search, for m pulses, times N multiply/add (assuming operation around 4.8kbits/sec)	5xN	1 band: 16x16=2N or 2 bands:6x2x16=1.5N or 3 bands:4x3x16=1.5N or 4 bands:3x4x16=1.5N Empty bands regeneration $\simeq$ N i.e. maximum total 3N
For synthesis (PxN)	12xN	3.5xN
Inverse filtering for gains	/	7xPxN=3.5N
Full search codebooks: For M level codebook of P dimensions the complexity is PxM For gains: (7 dimensional vector)	/	9 bit codebook = 512 levels 512x7=3584=28N
For LPC filters: (4-dimensional vectors)	/	bit distribution for 7 bands: 6 6 5 4 3 3 3 total number of levels: 64 + 64 + 32 + 16 + 9 + 9 + 9 =203 203x4=812 for every two frames=406/frame $\simeq$ 3.2N

The total number of multiply/add operations is 137N or 137 mult/add operations per sample for the multipulse whereas the equivalent total for the new algorithm is 76.2N.

In addition to the above the new algorithm possesses additional complexity due to the subband tree analysis-synthesis burden. For a 32-tap 3 stage QMF tree:

At each stage of the tree and for each output sample we have 16+16=32 multiply/add operations, or 16 multiply/add operations for every input sample. Therefore the complexity for the first stage is 16N. For the subsequent stages the input sampling rate is halved but the number of bands doubles. The complexity therefore remains constant at 16N. For a three stage analysis therefore the complexity is 3x16N=48N. The complexity for the synthesis is the same. Therefore the total complexity for the tree QMF analysis-synthesis is 96N. This brings the total complexity for the new algorithm to around 172N which is slightly higher than the simplified multipulse.

If one uses a filter with a smaller number of taps at each successive stage as it is normally the practice, or alternatively, if a parallel QMF structure is used it is possible to bring the complexity of the new scheme below that of the simplified multipulse algorithm.