Wideband (0-7kHz) speech coding techniques

This item was submitted to Loughborough University’s Institutional Repository by the/author.

Additional Information:

- A Doctoral Thesis. Submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of Loughborough University.

Metadata Record: https://dspace.lboro.ac.uk/2134/13785

Publisher: © Soo Ngee Koh

Please cite the published version.
This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (https://dspace.lboro.ac.uk/) under the following Creative Commons Licence conditions.

Attribution-NonCommercial-NoDerivs 2.5

You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

BY: Attribution. You must attribute the work in the manner specified by the author or licensor.

Noncommercial. You may not use this work for commercial purposes.

No Derivative Works. You may not alter, transform, or build upon this work.

For any reuse or distribution, you must make clear to others the license terms of this work.

Any of these conditions can be waived if you get permission from the copyright holder.

Your fair use and other rights are in no way affected by the above.

This is a human-readable summary of the Legal Code (the full license).

Disclaimer

For the full text of this licence, please go to: http://creativecommons.org/licenses/by-nc-nd/2.5/
WIDEBAND (0 - 7 kHz) SPEECH CODING TECHNIQUES

by

SOO NGEE KOH,

B.Eng. (University of Singapore, Singapore)
B.Sc. (University of London, England)
M.Sc. (Loughborough University of Technology, England)

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

October 1984

Supervisor: Constantinos S. Xydeas
Senior Lecturer, Electronic and Electrical Engineering
Department

© by Soo Ngee KOH, 1984
SYNOPSIS

With the existing telephone networks evolving towards Integrated Services Digital Networks, there is a specific need to transmit high quality wide-band (0-7 kHz) speech at 64 kbps or below for special services like voice channels in teleconferencing, commentary channels for broadcasting etc.

In this thesis, a computer simulation study of digital coding of wideband speech at 64 kbps using relatively simple coding techniques is first presented. The performance of ADPCM coders employing fixed or adaptive prediction, with or without noise spectral shaping, and 2-band subband coders is examined under ideal as well as noisy channel conditions. While preserving the quality of the 64 kbps recovered speech, the transmission bit rate is reduced to 56 kbps so that 8 kbps data can be accommodated within 64 kbps channel.

Channel capacity can be doubled if one could offer the same quality at 32 kbps. The first area of investigation at this bit rate is ADPCM employing adaptive noise spectral shaping. Though improved subjective quality is obtained compared to the conventional ADPCM systems without noise shaping, the recovered speech is found to be inadequate for wideband applications. Thus to achieve commentary-grade quality speech at this bit rate, sophisticated frequency-domain coders like SubBand Coding (SBC) and Adaptive Transform Coding (ATC), which exploit the time-varying properties of the speech signals and the auditory masking properties of the ear, are examined.

In subband coding, 7-band subband coders employing various combinations of fixed and adaptive bit allocation, forward and backward adaptive quantization and noise shaping techniques, are first studied. To fully exploit the advantage of adaptive bit allocation and to provide higher frequency resolution, the number of bands is increased to 14. In order to reduce the
complexity of the conventional bit allocation algorithm, a novel simplified bit allocation scheme is proposed. Its implementation is almost trivial and yet the quality of the recovered speech is found to be very close to that produced by the fully adaptive algorithm.

The use of Complex Quadrature Mirror Filters (CQMF) in a Complex SubBand Coder (CSBC) which divides speech into uniform channels of amplitude and phase signals is proposed as an alternative to the conventional split-band approach where QMF filter banks are used in the design of subband coders. A novel amplitude quantization strategy which exploits the correlations between the amplitude signals from the adjacent subbands is also proposed and used in conjunction with a simplified bit allocation algorithm for the quantization of the phase information. CSBC is found to be comparable to SBC in SNR measurements and subjective performance.

Finally, Adaptive Transform Coding of wideband speech is studied at 32 kbps. The quality of the recovered speech is found to be inferior to subband coding due to the block processing nature of ATC. Furthermore, complexity appears to be a serious limitation in implementing ATC coders. A reduced-complexity small blocksize pinned-sine transform speech coder is thus proposed. Its performance is found to be better than that of ADPCM systems in both SNR measurements and in informal subjective listening tests.
ACKNOWLEDGEMENTS

It gives me great pleasure to express my most sincere gratitude to my supervisor, Dr Costas Xydeas, for his continuous guidance, enormous encouragement and inspiration throughout the course of this research. Dr Xydeas’s invaluable knowledge in the field of speech coding, and his criticisms, helped to initiate and have greatly influenced the work presented in this thesis. Both his patience in reading and correcting the draft of the thesis and help in preparing the various publications are greatly appreciated also.

I wish also to thank Mr Chris Wheddon of British Telecom Research Laboratories for arranging the necessary financial support for the project. Many thanks are due also to Mr Fred Westall, Dr Roger Hanes and Mr Dominic Lee in R18.2.3 of B.T. for their interest in the work and their numerous helpful criticisms and suggestions. Their patience in the more than two dozen of informal subjective listening tests is greatly appreciated.

I am also grateful to Professor I.R. Smith, Head of the Electronic and Electrical Engineering Department, for providing the necessary research facilities. Many thanks are due also to my many colleagues in the communication lab for their assistance in many ways and their stimulating and thought provoking discussions. I am especially grateful to Dr F. Yeoh and Dr W.K. Cham for their helpful comments and suggestions.

Many thanks are due also to Mrs J. Brown for her forbearance in typing this thesis.
My deepest gratitude is due to my parents for their love and support and the sacrifices they have made throughout all the years of my education. I am greatly indebted also to my brothers and sisters for their concern and encouragement. Finally, I am deeply grateful to Miss Ong Siok Loh whose words of encouragement and unceasing moral support have helped me pass through many difficult moments in the past few years.
# GLOSSARY OF PRINCIPAL ABBREVIATIONS AND SYMBOLS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISDN</td>
<td>Integrated Services Digital Networks</td>
</tr>
<tr>
<td>IDN</td>
<td>Integrated Digital Networks</td>
</tr>
<tr>
<td>CCITT</td>
<td>Consultative Committee for International Telephone and Telegraph</td>
</tr>
<tr>
<td>PCM</td>
<td>Pulse Code Modulation</td>
</tr>
<tr>
<td>APCM</td>
<td>Adaptive PCM</td>
</tr>
<tr>
<td>DPCM</td>
<td>Differential PCM</td>
</tr>
<tr>
<td>ADPCM</td>
<td>Adaptive DPCM</td>
</tr>
<tr>
<td>ADM</td>
<td>Adaptive Delta Modulation</td>
</tr>
<tr>
<td>VPCM</td>
<td>Vector PCM</td>
</tr>
<tr>
<td>AQJ</td>
<td>Adaptive Jayant Quantizer</td>
</tr>
<tr>
<td>AQF</td>
<td>Adaptive Quantizer Forward</td>
</tr>
<tr>
<td>FBAP</td>
<td>Forward Block Adaptive Prediction</td>
</tr>
<tr>
<td>BSAP</td>
<td>Backward Sequential Adaptive Prediction</td>
</tr>
<tr>
<td>SAP</td>
<td>Stochastic Approximation Prediction</td>
</tr>
<tr>
<td>SGEF</td>
<td>Sequential Gradient Estimation Prediction</td>
</tr>
<tr>
<td>SNRSEG</td>
<td>Segmental Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>PDE</td>
<td>Pre-emphasis and De-emphasis</td>
</tr>
<tr>
<td>PPF</td>
<td>Pre-filtering and Post-filtering</td>
</tr>
<tr>
<td>ANS</td>
<td>Adaptive Noise Shaping</td>
</tr>
<tr>
<td>LPT</td>
<td>Low-Pass Translation</td>
</tr>
<tr>
<td>BPT</td>
<td>Band-Pass Translation</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite Impulse Response</td>
</tr>
<tr>
<td>IIR</td>
<td>Infinite Impulse Response</td>
</tr>
<tr>
<td>QMF</td>
<td>Quadrature Mirror Filter</td>
</tr>
</tbody>
</table>
CQMF : Complex Quadrature Mirror Filter
ABA : Adaptive Bit Allocation
FBA : Fixed Bit Allocation
SBA : Simplified Bit Allocation
IBAP : Invariant Bit Allocation Pattern
SBC : Subband Coding
CSBC : Complex Subband Coding
PARCOR : Partial Correlation
pdf : Probability Density Function
psd : Power Spectral Density
BER : Bit Error Rate
ATC : Adaptive Transform Coding
RBC : Recursive Block Coding
DFT : Discrete Fourier Transform
IDFT : Inverse DFT
KLT : Karhunen-Loeve Transform
PKLT : Pinned-KLT
DCT : Discrete Cosine Transform
DST : Discrete Sine Transform or Discrete Slant Transform
PST : Pinned-Sine Transform
DSP : Digital Signal Processing
\( x(n), x_n \) : input speech sample
\( \hat{x}(n), \hat{x}_n \) : recovered speech sample
\( a_k \) : prediction coefficient
A : prediction coefficient vector
\( \beta, \gamma \) : leakage constant or noise shaping parameter
\( \Delta \): quantizer step-size

\( G_{TC} \): transform coding gain over PCM

\( w(n) \): data window sample

\( f_m(n) \): mth forward residue sample at time instant n

\( b_m(n) \): mth backward residue sample at time instant n

\( f^s_m(n), b^s_m(n) \): sign of \( f_m(n) \) and \( b_m(n) \) respectively

\( k_m(n) \): mth-stage reflection coefficient at time instant n

\( M(.) \): time-invariant multiplier function

\( E[x] \): expectation of x

\( \sigma \): standard deviation

\( \sigma^2 \): variance

\( \langle \cdot \rangle \): average value

\( I \): identity matrix

\( Q^{-1} \): inverse of square matrix Q

\( Q^T \): transpose of matrix Q

\( \delta_{ij} \): Kroneker delta function

\( f_s \): sampling frequency

\( \omega \): angular frequency

\( f_c \): cut-off frequency

\( T \): sampling period

\( R_i \): number of bits allocated to the ith frequency component or transform coefficient

\( \overline{R} \): average bit rate

\( X(\omega) \): Fourier transform of \( x(n) \)

\( * \): convolution
CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>INTRODUCTION</th>
<th>SPEECH COMMUNICATION AND TELEPHONY</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>SPEECH COMMUNICATION</td>
<td>DIGITAL SPEECH COMMUNICATION</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>APPLICATION OF WIDEBAND SPEECH AND THE INTEGRATED SERVICES DIGITAL NETWORKS (ISDN)</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1.3.1</td>
<td>Application of Wideband Speech</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>1.3.2</td>
<td>The ISDN</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>1.4</td>
<td>ORGANISATION OF THESIS</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>SPEECH PRODUCTION AND CODING</th>
<th>THE PRODUCTION OF SPEECH SOUNDS</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>THE PRODUCTION OF SPEECH SOUNDS</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>2.2</td>
<td>CHARACTERISTICS OF SPEECH SIGNALS</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>2.3</td>
<td>CLASSIFICATION OF SPEECH CODING SYSTEMS</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>2.3.1</td>
<td>Analysis-Synthesis Systems</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>2.3.1.1</td>
<td>Channel vocoder</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>2.3.1.2</td>
<td>Formant vocoder</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>2.3.1.3</td>
<td>Linear prediction coding (LPC) vocoder</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>2.3.2</td>
<td>Time-Domain Coder</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>2.3.2.1</td>
<td>Pulse Code Modulation (PCM)</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(i) Uniform quantization</td>
<td>43</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(ii) Non-uniform quantization</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(iii) Adaptive quantization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2.3.2.2 Vector Pulse Code Modulation (VPCM)  
2.3.2.3 Differential Pulse Code Modulation (DPCM)  
   (i) Adaptive DPCM (ADPCM)  
   (ii) Adaptive prediction  
      (a) Forward Block Adaptive Prediction (FBAP)  
      (b) Backward Sequential Adaptive Prediction (BSAP)  
      (b.1) Stochastic Approximation Prediction (SAP)  
      (b.2) Sequential Gradient Estimation Prediction (SGEP)  
2.3.2.4 Delta Modulation (DM)  
   (i) Adaptive Delta Modulation (ADM)  
      (a) Constant Factor Delta Modulation (CFDM)  
      (b) Continuously Variable Slope DM (CVSD)  
2.3.3 Frequency-Domain Coder  
   2.3.3.1 Subband Coding (SBC)  
   2.3.3.2 Adaptive Transform Coding (ATC)  
   2.3.3.3 Phase Vocoder ($\phi$V)  
2.3.4 Hybrid Coding  
   2.3.4.1 Residual-Excited Linear Prediction (RELP) Vocoder  
      (i) Spectral flattening  
   2.3.4.2 Frequency Scaling of Speech Signals  
      (i) Analytic Signal Rooting Technique  
      (ii) Phase Vocoder
(vi)

(iii) Frequency-Domain Harmonic Scaling (FDHS) 94
(iv) Time-Domain Harmonic Scaling (TDHS) 95
(v) The TDHS/(SBC, ATC) speech coders 98

2.4 ASSESSMENT OF CODER PERFORMANCE 98
2.4.1 Segmental Signal-to-Quantization Noise Ratio (SNRSEG) 101
2.4.2 Long-Term Average Power Spectral Density of Output Noise 102
2.4.3 Informal Subjective Listening Tests 103

2.5 HARDWARE ISSUES AND CODER COMPLEXITY 104
2.5.1 Hardware Issues 104
2.5.2 Coder Complexity 105

2.6 CONCLUSIONS 108

CHAPTER 3 TIME AND FREQUENCY DOMAIN CODING OF WIDEBAND SPEECH AT 64 AND 56 KBPS 109

3.1 INTRODUCTION 109
3.2 ADPCM EMPLOYING FIXED PREDICTION 113
3.2.1 Adaptive Jayant's Quantizer 117
3.2.1.1 Robust Jayant's Quantizer 119
3.2.2 Determination of Fixed Prediction Coefficients 121
3.2.2.1 Autocorrelation Method 121
3.2.2.2 Covariance Method 122
3.2.2.3 Long-Term Fixed Prediction Coefficients for Wideband Speech Sampled at 16 kHz 123
(vii)

3.3 ADPCM EMPLOYING ADAPTIVE LATTICE PREDICTION 125
   3.3.1 Adaptive Lattice Prediction 126
   3.3.2 The Sign Product Method 132

3.4 ADPCM EMPLOYING SIMPLE NOISE SPECTRAL SHAPING 134

3.5 TWO-BAND SUBBAND CODING WITH ADPCM ENCODERS 138
   3.5.1 Two-Band SBC Coder with Noise Spectral Shaping 140

3.6 RESULTS AND DISCUSSIONS 143
   3.6.1 ADPCM/AQJ with Fixed First/Second Order Prediction 143
   3.6.2 ADPCM/AQJ with Adaptive Lattice Prediction 145
   3.6.3 ADPCM/AQJ with Simple Noise Spectral Shaping 148
   3.6.4 Two-Band Subband Coding with ADPCM coders 150
   3.6.5 μ-Law PCM at 128 kbps 153
   3.6.6 Final Remarks 155

3.7 PERFORMANCE OF THE ADPCM/LATTICE AND SBC CODERS UNDER TRANSMISSION ERROR CONDITIONS 157
   3.7.1 Performance of the ADPCM/Lattice Coder Under Transmission Error Conditions 158
   3.7.2 Performance of the Two-Band SBC Coder Under Transmission Error Conditions 161

3.8 REAL-TIME IMPLEMENTATION OF A TWO-BAND SUBBAND CODER 164

3.9 PERFORMANCE OF A TWO-BAND SUBBAND CODER AT 56 KBPS 168

3.10 WIDEBAND SPEECH CODING RESEARCH WITHIN THE CCITT 171
CHAPTER 4  ADPCM CODING OF SPEECH WITH ADAPTIVE NOISE SPECTRAL SHAPING

4.1 INTRODUCTION

4.2 ADAPTIVE PREDICTIVE CODING (APC) OF SPEECH

4.3 ADAPTIVE NOISE SPECTRAL SHAPING (ANS) TECHNIQUES

4.3.1 Adaptive Noise Spectral Shaping by Atal and Schroeder

4.3.2 Adaptive Noise Spectral Shaping by Makhoul and Berouti

4.3.3 Comparison Between the ANS Techniques Proposed by Atal and Makhoul

4.3.4 Adaptive Noise Spectral Shaping via the Forward Adaptive Pre- and Post-Filtering Technique

4.3.5 Backward Adaptive Noise Spectral Shaping

4.4 ADPCM CODING OF WIDEBAND SPEECH WITH AND WITHOUT NOISE SHAPING AT 32+ KBPS

4.5 RESULTS AND DISCUSSIONS

4.5.1 Results and Comparisons of the ADPCM and NS1 Systems

4.5.2 Results of the NS2 System

4.5.3 Comparisons of the ADPCM, NS1 and NS2 Systems

4.5.4 Results of the NS3 System

4.5.5 Results of the NS4 System
5.1 INTRODUCTION

5.2 SPEECH CODING IN SUBBANDS

5.2.1 Division of Speech Spectrum into Subbands
5.2.2 Low Pass Translation (LPT) of Subband Signals
5.2.3 Quadrature Mirror Filter (QMF)
5.2.4 Complex Quadrature Mirror Filter (CQMF)
5.2.5 Fixed and Adaptive Bit Allocation (FBA/ABA)
5.2.6 Simplified Bit Allocation Algorithm (SBA)
5.2.7 Vector Quantization of Bit Allocation Patterns

5.3 SEVEN-BAND SUBBAND CODERS (7/SBC)

5.3.1 Seven-Band Subband Coding with Fixed Bit Allocation and Uniform AQFs
5.3.1.1 Results and Discussions
5.3.2 Seven-Band Subband Coding with Fixed Bit Allocation and ADPCM/AQJ Encoders
5.3.2.1 Results and Discussions
5.3.4 Seven-Band Subband Coding with Simplified Bit Allocation and Laplacian AQFs
5.3.5 Seven-Band Subband Coding with Adaptive Pre- and Post-Filtering, Fixed Bit Allocation and AQFs
5.3.5.1 Results and Discussions
5.3.6 Summary of the Performance of the 7/SBC coders
5.4 FOURTEEN-BAND SUBBAND CODERS (14/SBC)  
5.4.1 Fourteen-Band Subband Coding with Fixed Bit Allocation, ADPCM/AQJ and AQFs  
5.4.1.1 Results and Discussions  
5.4.2 Fourteen-Band Subband Coding Employing the ABA or SBA Algorithm and Laplacian AQFs  
5.4.2.1 Results and Discussions  
5.4.3 Comparison Between the 14/SBC/AQF Coding of Wideband and Narrowband Speech at the Bit Rate of 16 kbps  
5.4.4 Comparison Between the 14/SBC/AQF coder at 32 kbps and the 2-band SBC coder at 64 kbps  

5.5 SEVEN-BAND COMPLEX SUBBAND CODERS (7/CSBC)  
5.5.1 Seven-Band Complex Subband Coding with Simplified Bit Allocation and AQFs  
5.1.1 Results and Discussions  
5.5.2 Seven-Band Complex Subband Coding with Seven Fixed Bit Allocation Patterns and a Novel Amplitude Quantization Technique  
5.5.2.1 Results and Discussions  

5.6 NOTE ON PUBLICATION  
5.7 SUMMARY AND CONCLUSION
CHAPTER 6  ADAPTIVE TRANSFORM CODING (ATC) OF SPEECH SIGNALS

6.1 INTRODUCTION  335

6.2 BASIC THEORY AND ISSUES RELATED TO ATC  338
  6.2.1 The Block Transformation  338
  6.2.2 Quantization of the Transform Coefficients  343
  6.2.3 Determination of the Basis Spectrum  345

6.3 THE PINNED-KL AND PINNED SINE TRANSFORM  348
  6.3.1 Signal Representation  350
  6.3.2 The Pinned-KL Transform (PKLT)  351
  6.3.3 Recursive Block Coding Technique  355
  6.3.4 The Pinned-Sine Transform (PST)  357

6.4 LARGE BLOCKSIZE ATC WIDEBAND SPEECH CODER AT 32 KBPS  359
  6.4.1 Smoothing Techniques to Reduce the Effect of
         Interblock Discontinuities  360
  6.4.2 Results and Discussions  363

6.5 SMALL BLOCKSIZE ATC WIDEBAND SPEECH CODERS EMPLOYING THE
       PINNED-KL OR PINNED-SINE TRANSFORM  366
  6.5.1 Results and Discussions  372
  6.5.2 Comparison Between the 32 kbps Small Blocksize
       PST and Large Blocksize ATC coders with the
       64 kbps 2-Band SBC Coder  375

6.6 COMPARISONS OF THE 32 KBPS TIME AND FREQUENCY-DOMAIN
       WAVEFORM CODERS AND THE 64 KBPS 2-BAND SBC CODER  378

6.7 SUMMARY AND CONCLUSION  385
CHAPTER 7  RECAPITULATION AND CONCLUSION

7.1  RECAPITULATION  387
7.2  SUGGESTIONS FOR FUTURE RESEARCH  394
7.3  CLOSING REMARKS  396

APPENDIX I TO APPENDIX VI

REFERENCES
APPENDICES

(I) : Relation Between the Cross-Correlation Function of Two Random Variables and that of the Signs of the Two Random Variables.

(II) : Proof of the Constraint on Quantization Noise Spectrum.

(III) : Equations Relating to Complex Quadrature Mirror Filter.

(IV) : Input Decision Thresholds and Output Levels of Gaussian and Laplacian Quantizer.

(V) : The Correlation Properties of the $v_i$ sequence.

(VI) : Main Processing and Storage Requirements of the Various Frequency-Domain Waveform Coders.
CHAPTER I

INTRODUCTION
INTRODUCTION

1.1 SPEECH COMMUNICATION AND TELEPHONY

The development of human civilization is made possible by man's ability to share experience, to exchange ideas and to pass down knowledge from one generation to another through vocal and written communications. Though carved symbols and ideograms on the stone walls of caves, clay tablets and animal bones or cells found among the relics of the primitive societies, and the invention of the printing press, paralleled the development of society, it is undeniably true that voice communication is by far the most important and convenient means of human communication. Indeed, the first cry of every new-born baby signals the beginning of vocal communication between the baby and his environment. With this inherent ability to produce basic utterances like crying, laughing and groaning, he develops to a very sophisticated system of expressing his wishes and feelings through the use of complicated, but systematic, sounds produced by his vocal organ. Collectively, these different sounds are termed 'speech'. Speech is closely related to his ability to think abstractly and it is also one of those few basic abilities that sets him apart from animals.

Speech is not a mere concatenation of different phonemes (sounds). It involves virtually continuous movements of the speech production organ and is complicated and enriched by the subtle use of inflection, pitch,
loudness and intonation. A very simple word, 'yes', said in different ways by applying different pitch, loudness, length and inflection, carries quite different meanings. In a tone language like Chinese, pitch variations achieve the function of distinguishing words with different meanings. Thus in modern standard Chinese, the same sequence of phonemes /ma/ can represent four different words according to the pitch, loudness and length assigned to the syllable: high level pitch /-ma/ means mother; high rising pitch /\ma/ means insensitive; low, long fall-rise pitch /\ma/ means a horse; and high falling pitch /\ma/ means to curse. Take another example - try to talk to a cat or dog using inflection, pitch and loudness, but with meaningless words. The animal will respond. Then speak to it in monotone, but using meaningful words. Let the voice drone on and on and you can actually make the animal fall asleep.

By applying phonetic rules, sequences of phonemes can be formed into words. Words are then organised into phrases and phrases are assembled into sentences according to grammatic and semantic rules. Of course human speech is not simply mechanically formed phrases and sentences. It expresses feelings and emotion, indicates intellectual prowess and reflects life's experiences. It can be used to encourage or intimidate, anger or pacify, soothe or excite, command or persuade. Indeed, many successful people are noted for their masterful communication skills.
Not until man knew how to harness the power of electric current did he develop the ability to satisfy partially his desire for rapid communication over long distances between any two points. In 1858, twenty six years after Morse invented the telegraph, a submarine telegraph cable was established between North America and Great Britain\(^{(1)}\). It was no more than a system which accomplished the conveyance of written messages at the speed of light. Though quite a milestone in the development of human civilization, its impact was not revolutionary and its use restrictive.

In 1876, Alexander Graham Bell invented the telephone and started the era of voice communication over long distances. Basically, the telephone instrument consists of two separate transducers and some supporting electrical components. One of the transducers converts the acoustic signal of human speech into an electrical signal which is transmitted to the called party at the speed of light. Conversely, the other transducer produces an acoustic signal upon the reception of an electrical signal from another instrument. With a number of subsequent technical developments like high-frequency transmission, single sideband suppressed carrier and steerable antennas, overseas radio telephony was introduced in 1927\(^{(1)}\). The inventions and subsequent developments of the electron tube and the negative feedback principle enhanced the design of the deep-water repeater and it was not until 1956 that the first transatlantic submarine cable, for voice communication between North America and Great Britain, went into service\(^{(1)}\).
The invention of the telephone brought with it a host of related technologies like switching and transmission technologies. By 1982, there were 550 million telephones on the earth\(^{(2)}\). Voice communication network has become one of the basic elements in the infrastructure of all the developed, and most of the developing, societies. It has become a necessity in life and ensures the proper functioning of any civilized society. In the future, the development of integrated circuits and advancement in digital processing and computer technologies will enhance and improve the quality of voice communication over the gradually emerging Integrated Services Digital Networks (ISDN).
1.2 DIGITAL SPEECH COMMUNICATION

In general, the existing telecommunication systems can be divided into the analogue and digital systems. With the achievement of low cost large-scale integration and the development of digital technologies, digital communication will gradually dominate the future network. Figure 1.1.1 shows a block diagram of a basic digital communication system (DCS). The function of the source coder is to convert an analogue audio or video signal $x(t)$ into binary digits $s_i$. Channel codec is usually added to mitigate the deleterious effect introduced by the imperfect transmission channel. With proper system design and normal channel condition, $s_i = \hat{s}_i$. However, $x(t)$ is not necessarily equal to $\hat{x}(t)$ even under the equality of $s_i$ and $\hat{s}_i$. The source coder is usually designed in such a way that $x(t)$ can be represented by the minimum number of bits and yet the recovered $\hat{x}(t)$ is perceived as $x(t)$ within an acceptable subjective criterion. Indeed, the main part of the rest of this thesis will discourse on the different designs of source coders for the coding of speech signals at various bit rates.

The full scale modernization of our communication systems into digital ones is not without reasons. One can be easily convinced of the necessity to go into digital by the main advantages and reasons listed below:

(1) In analogue communication systems, the transmission channels have to be linear since non-linearity causes signal distortion.
Figure 1.1.1  Basic digital transmission system
They also tend to accumulate noise and other impairments with distance. In contrast, information in digital communication systems can be transmitted over long distances without any degradation at all because digital signals can be retimed and reshaped along the transmission path.

(2) The availability of a large range of digital processing and storage devices at continually dropping cost means that in most areas digital operations are more cost effective than analogue operations.

(3) Digital equipments are free from the effects of drift and the need for complicated setting up and adjustment. Maintenance is thus simplified.

(4) Many processes impracticable or impossible using analogue techniques can readily be implemented using digital technologies. This applies particularly to processes relying on signal storage. Signal processing techniques like bit-rate reduction and encryption are now amenable using digital signal processing technologies.

(5) The arrival of the fifth and further generation computers and the full exploitation of information technologies in the near future will bring an immense increase in man-machine and machine-machine communications. By then, digital communication of signals like speech and video will be unavoidable.
It is worth noting that young engineers these days generally emerge from college or university with a much better grasp of digital than analogue techniques. With the current emphasis in computer education for all ages of students, it would be a much easier task to train a work force to deal with digital systems and thus increase productivity.

Since the invention of the first digital waveform coding technique, Pulse Code Modulation (PCM), of speech signal by Reeves in 1938, an enormous amount of different algorithms based on different design concepts which exploit the various time and frequency-domain properties of speech signals have evolved. The underlying goal of all the coder designers is not new. It is to transmit speech with the highest possible quality, using the least number of bits and yet at a reasonable cost. These requirements are unfortunately self-conflicting. The tasks of the design engineer is to find a compromise and gear a specific design for a specific application. For example, high quality speech and cheap terminal equipment are of higher priority than channel capacity for ordinary voice communication over the public networks. On the other hand, voice communication using the least channel capacity is a more important factor than quality and equipment cost for military application.
1.3 APPLICATION OF WIDEBAND SPEECH AND THE INTEGRATED SERVICES DIGITAL NETWORKS (ISDN)

1.3.1 Application of Wideband Speech

The existing telephone channels restrict the bandwidth of speech signals to within the 0.3 - 3.4 kHz frequency range and employ log-PCM for digital coding at 64 kbps (4). As the frequency range of unvoiced sounds like 's', 'sh', 'f' or 'th' extends beyond the 3.4 kHz limit, this bandwidth limitation causes attenuation and distortion resulting in loss of speech intelligibility and quality. If the bandwidth of the communication channels can be increased to the frequency range of 0 - 7 kHz, substantial improvements in the fidelity of the transmitted speech can be achieved (5). Speech signals having a bandwidth of 7 kHz are usually termed wideband or commentary grade signals.

With many national communication networks evolving towards digital end to end connections, there is increasing interest within telecommunications administrations in providing high quality voice services using new coding techniques more efficient than PCM. Digital encoding of narrowband speech at 32 kbps and wideband speech at 64 kbps in the Integrated Services Digital Networks (ISDN) are targeted for implementations in the near future (6). Possible applications of wideband speech for better quality voice services include loudspeaker telephones, voice channels for audio and video conferencing and commentary channels.
in broadcasting. The goal of this research is to study various digital coding techniques when encoding wideband speech at the bit rate of 64 kbps and below.
1.3.2 The Integrated Services Digital Networks (ISDN)

The ever growing proliferation of different types of digital networks, e.g. the circuit switched Integrated Digital Network (IDN) for telephony and bulk data and the packet switched IDN for bursty data, imposes burdens on both network users and providers. Together with the rapid growth in computer technologies and the unlimited demands for communication facilities, they stimulated the emergence of the concept of ISDN.(7) An ISDN is characterised by three main features, namely (1) end-to-end digital connectivity, (2) multiservices capability (voice, video and data) and (3) standard terminal interfaces(8). The conversion of the existing network to the final comprehensive ISDN may require a long period of time. During the transition period new network and system architectures are being introduced to partially meet the essential elements of the ISDN concept definition:

Network Architectures (7-9)

The early stage of ISDN will provide narrowband services like voice and data transmission. It is characterised by the hybrid of circuit and packet switching capability. Bulk data and voice are treated by the 64 kbps circuit switching and common channel signalling while bursty data are handled by packet switching. Figure 1.3.2 illustrates the introductory phase of an ISDN. Two types of communication channels are defined. The B-channel operates at 64 kbps. It is a circuit switched channel for voice or data. The D-channel consists of an
8 kbps circuit switched channel for data and another 8 kbps channel for message based signalling. The basic access structure consists of a TDM assembly of one or two B-channels and one D-channel.

Typical uses of B-channels include the transmission of (1) PCM narrowband speech at 64 kbps or wideband speech at the same bit rate using more efficient digital encoding techniques, (2) data information corresponding to circuit switching or packet switching data user classes of bit rates adapted to 64 kbps and (3) low bit rate voice (LBRV) combined with data information. More advanced ISDN will provide enhanced voice and data transmission and further include wideband services like wideband video via new fiber distribution plants and local ISDN exchanges appended with wideband switching capabilities. Figure 1.3.3 shows the phase 2 ISDN concept. Of course the uses of the ISDN channels are by no means unchangeable. The 64 kbps wideband speech channel can be alternatively adapted for wideband speech at 56 or 48 kbps with the remaining 8 or 16 kbps capacity for low speed data.

All public network providers are moving aggressively towards ISDN implementation. Transition services are targeted for 1984-1985 while the full ISDN services are expected to be a reality in 1987. The success of ISDN will contribute to greater customer satisfaction and prepare a country to meet the challenges in the age of Information Technology.
**Figure 1.3.1** A reference configuration for the ISDN

**Figure 1.3.2** Early ISDN architecture for voice and data capability (after Reference 8)

**Figure 1.3.3** Advanced ISDN architecture with wideband capability (after Reference 8)
1.4 ORGANISATION OF THESIS

This thesis starts with a general discussion on speech communication and an outline of telephony. A description of a digital communication system and the advantages of transmission of speech using digital techniques are presented next. Following that, the application of wideband speech, which is the motivation of the research, and the general configuration of the forthcoming Integrated Services Digital Networks, are discussed.

The second chapter mainly describes the general speech coding techniques. Before that, the basic structure of the human voice production system and the mechanism of speech production are outlined. The basic time and frequency-domain characteristics of the speech signals are presented followed by an introduction about the three categories of speech coding techniques, namely the Analysis-Synthesis or Vocoding techniques, waveform coding in both time and frequency domain and the hybrid coding techniques. The coverage of each of these categories of techniques is wide-ranging, rather than in-depth, and with the minimum amount of mathematics. However, essential information is given to give the reader a basic understanding of the principles behind the techniques. The chapter ends with a description of the performance criteria used in assessing the various coding techniques proposed and studied in this research and a brief discussion on the issue of coder complexity.
In Chapters 3 through 6, the research work conducted is presented. Chapter 3 presents a comparative study of the various time and frequency-domain waveform coding techniques applied to wideband speech at 64 kbps. It ends with a brief discussion on the recent activities within CCITT on the study of wideband speech coding techniques. In Chapter 4, the application of adaptive noise spectral shaping in coarse quantization ADPCM coding of wideband speech is presented. The adaptive noise spectral shaping techniques investigated are the techniques of adaptive noise feedback filtering and the adaptive pre- and post-filtering technique. The use of fixed first order pre-emphasis as the simplest form of noise spectral shaping is also discussed and compared with other adaptive noise shaping techniques. Chapter 5 presents a detailed investigation on subband coding of wideband speech at 32 kbps. Various 7 and 14-band subband coding schemes employing different bit allocation and quantization strategies are discussed. A novel simplified bit allocation algorithm is described and its performance, when applied to the subband coders, is compared with that of the conventional bit allocation algorithm. The last part of Chapter 5 is devoted to the description of the design of a complex subband coder (CSCB) based on the less known complex Quadrature Mirror Filter (CQMF). A novel amplitude quantization strategy which exploits the correlation between the amplitude samples from adjacent subbands is proposed and used for the coding of the amplitude signals. The performance of the subband coders are assessed and compared in terms of average segmental SNR measurements, power spectral density plots.
of the output noise of the coders and informal subjective listening
tests. In Chapter 6, the performance of the conventional large block-
size adaptive transform coding (ATC) of wideband speech at 32 kbps is
examined. The problem of inter-block discontinuities arising from
the block processing nature of an ATC coder at low bit rate is high-
lighted and a solution is proposed. To avoid the complexity problem
associated with large blocksize ATC coder, the use of a small blocksize
pinned-sine transform coder for the coding of wideband speech at
32 kbps is considered and its performance compared with that of the
conventional ATC coders. This chapter ends with a comparison of the
computational complexity of the frequency-domain 32 kbps wideband speech
coders considered so far in the research.

The final chapter provides a recapitulation of the work done in the
research, the novel techniques proposed and the results obtained.
Suggestions are made for further research along the directions already
investigated and finally, a conclusion is made.
CHAPTER 2

SPEECH PRODUCTION AND CODING
2.1 THE PRODUCTION OF SPEECH SOUNDS \(^{(10,11)}\)

Acoustically, the human voice production system consists of an excitation source and a time-varying resonant cavity formed by the vocal tract. Figure 2.1.1 shows the cross-sectional view of the human vocal tract.

The source of excitation for voice production is the stream of air that comes from the lungs as we exhale. It passes through the glottis, which is the opening between the two vocal cords, and enters the vocal tract. The vocal tract is a non-uniform tube about 17 cm in length. It is terminated at one end by the vocal cords and at the other end by the lips. The cross-sectional area of the tract is determined by the movement of the velum, tongue, jaw and lips. The nasal tract forms an ancillary cavity to the main resonant cavity. It begins at the velum and terminates at the nostrils. For the production of nasal sounds, the nasal cavity is coupled to the main cavity by releasing the velum. Otherwise, the velum seals off the nasal tract and no air is radiated from the nostrils.

Sounds can be generated in the vocal system in three basic ways. For voiced sounds, the excitation source is a periodic train of air puffs. Air pressure builds up behind the vocal cords barrier and eventually blows the cord apart. The excess pressure is thus released and the cords return to their closed positions. The
pressure builds up again and the cycle repeats. Thus the vibrating vocal cords open and close the air passage between the lungs and mouth periodically. The steady flow of air is therefore transformed into a train of air puffs. The frequency of this air puff is determined by the lengths and the tension of the vocal cords. Subjectively, it is related to the pitch of the sounds one perceived. When the air puffs pass through the vocal tract, its frequency spectrum is modified by the different shape of the vocal tract. Thus different sounds are produced. The vowel sounds like "a", "e", "i", "o", "u" are produced in this way. They are collectively called voiced speech.

Unvoiced speech or fricative sounds of speech are generated by forming a constriction at some point in the tract and forcing air through it to produce turbulence. The source of sound is therefore noise or hiss-like. Speech sounds like "s", "sh", "f" and "th" belong to this category.

Plosive sounds result from making a complete closure by blocking the vocal tract with the tongue or the lips, and then suddenly releasing the air pressure built up behind the block. We use this "blocking" technique to produce sounds like "p", "g" and "t". It should be noted that fricative or plosive sounds are produced independent of the vocal cord activity. Whichever of the three techniques is used, the vocal tract modifies the frequency spectrum of the excitation source. The sound production mechanism therefore can be approximated by the rather simple linear system shown in Figure 2.1.2.
Figure 2.1.1  Cross-sectional view of the vocal tract
(after Reference 10)

Figure 2.1.2  Model of speech production
Though the source-tract interaction is still unknown, they are generally assumed to be linearly separable\(^{(11)}\). Figure 2.1.3 shows the source-tract model of speech production with underlying linear system theories. The sound \(s(t)\) radiated from the mouth can be approximated as the convolution of the excitation source \(g(t)\) and the vocal tract impulse response \(h(t)\), i.e.

\[
s(t) = g(t) \ast h(t)
\]  

\(2.1.1\)

where \(\ast\) denotes convolution. It must be borne in mind that \(h(t)\) is time-varying and so is \(g(t)\) in the long-term. However, they can be considered as stationary in the short-term. This basic concept is the basis of most vocoders to be discussed later.
Figure 2.1.3 Source-tract model of speech production
(after Reference 11)
2.2 CHARACTERISTICS OF SPEECH SIGNALS

Speech signals exhibit considerable redundancy because of the physical mechanism of their production. For voiced speech, the excitation source itself is periodic. The vocal tract is an acoustic resonator which usually has three dominant resonant frequencies below the frequency of 4 kHz. The time waveforms and the 3-dimensional squares of the magnitudes of the frequency components of speech are shown in Figures 2.1.4 and 2.1.5 respectively.

For voiced sounds, we observe in the time waveforms a pseudo-periodic structure which is reflected in the frequency spectrum, shown in Figure 2.1.6, as a harmonic structure in frequency. The envelope of the spectrum is determined primarily by the frequency response of the vocal tract filter. The resonances of this filter are observed as the peaks in the spectral envelope. They are referred to as formants. As compared to the time waveforms, they move about more regularly and consistently.

For unvoiced speech, both the time waveforms and the fine structure of the frequency spectrum are noise-like. Its spectral envelope also exhibits peak(s), for certain unvoiced sounds, in the region of 4 to 6 kHz, due to the resonance of the tract between the place of constriction, i.e. the tongue and the upper teeth and the opening at the lips.
The time waveforms and frequency plots in Figure 2.1.4 and Figure 2.1.6 show that speech is nonstationary in the long-term though there is a short-term stationarity within individual voiced and unvoiced segments. To understand its properties in the long-term, one can plot its long-term average spectral density (12) as shown in Figure 2.1.7. It is clear that high frequency signal components (above 3.5 kHz) contribute very little to the total speech energy. Due to this reason, the existing telephone channels limit the bandwidth to below 3.4 kHz so that channel capacity can be fully utilized. This bandlimited speech is commonly known as narrowband speech. Its intelligibility is about 85% (13). Though universally accepted as adequate for ordinary voice communication purposes, its quality leaves much to be desired. One can certainly recall that on many occasions, it is necessary to use words to represent letters, like 'c for Charlie', in order to get the message across the telephone channel. Thus information is not efficiently transmitted.

In order to provide a higher quality voice service over the forthcoming ISDN, the consultative committee for International Telephone and Telegraph (CCITT) has decided to include the transmission of wideband (0 - 7 kHz) speech signals as an additional voice service in the design of the future networks. Wideband speech is chosen because it has an almost complete intelligibility of 99% (13).
Figure 2.1.4a  Time waveform of 0.16 sec of voiced speech

Figure 2.1.4b  Time waveform of 0.16 sec of unvoiced speech
Figure 2.1.5a  3-D magnitude squares plots of 0.16 sec of voiced speech

Figure 2.1.5b  3-D magnitude squares plots of 0.144 sec of unvoiced speech
Figure 2.1.6 Amplitude spectra of 12 msec of
(a) voiced speech and
(b) unvoiced speech
Figure 2.1.7 Long-term average spectral density of speech (after Reference 12)
2.3 CLASSIFICATION OF SPEECH CODING SYSTEMS

Speech coding techniques can be broadly classified into three categories, namely analysis-synthesis (vocoder) coding, waveform coding and hybrid coding. The concepts used in the first two methods are very different. The third category of speech coding systems combines the various techniques used in the other categories to either achieve a further bit rate reduction or to improve the quality of the decoded speech.

In the analysis-synthesis systems the speech production mechanism is represented by a simplified theoretical model. The parameters of the model are derived from the actual speech signals at the transmitter. They are digitally encoded and transmitted to the receiver. At the receiver, the decoded parameters are used to control a speech synthesizer which corresponds to the model used in the analyser. The analysis-synthesis coders can achieve great bit rate reduction because it is speech specific. Both transmitter and receiver have the prior knowledge of the speech production model. Secondly, only perceptually significant parameters of the speech production model are extracted and transmitted. The synthesized speech perceived resembles the original speech but no effort is made to produce a close replica of the original speech waveform. The drawback of this category of speech coders is the complex processing
operations required. Examples of vocoders include the channel vocoder, formant vocoder, homomorphic vocoder and the linear predictive coding (LPC) vocoder.

In waveform coding systems, the number of bits available are made use of in the most efficient way to encode the speech signals in the time or frequency-domain so that the closest possible replica of the original speech waveform can be reproduced at the receiver. To achieve bit rate reduction, the redundancies inherent in speech signals are removed by this kind of coder. More recently, the perceptual effect of the quantization noise in relation to the speech spectrum is exploited to achieve an improved subjective quality for the recovered speech. In general, this category of coders are the least demanding in terms of implementation complexity. However, they operate at the highest bit rate among the three classes of coders.

The last category of speech coding system, i.e. the hybrid coding system, combines the various techniques employed in the vocoding and waveform coding systems. An example of this is the residual excited LPC (RELP). In this system, an all pole model of speech production is assumed. The parameters of the all pole filter are encoded in the usual way employed in a vocoding system. The residual signal at the output of the model is also encoded in the time-domain and transmitted to the receiver. Furthermore, the decoded residual signal is used as an excitation source for the same speech production model at
the receiver. No attempt is made to represent the excitation source as periodic pulse trains for voiced speech and random noise for unvoiced speech as in the case of the LPC vocoder. In another example, harmonic scaling in the time or frequency domain to reduce the bandwidth of the input signals is performed followed by waveform coding to achieve bit rate reduction. As can be expected, the complexity of this type of coder is in between that of the vocoding and waveform coding systems. The bit rates and quality of the three categories of speech coding systems are summarized in Table 2.3.1.

2.3.1 Analysis-Synthesis Systems

Digital speech coders employing analysis-synthesis techniques are commonly known as vocoders (contraction of the words voice and coder). These coders are characterised by (1) the presumption of a reasonably accurate speech production model and (2) the separation of the excitation and vocal tract information. A generalized block diagram of vocoders is shown in Figure 2.3.1.

Based on a specific speech production model, the analyser at the transmitter extracts information about the excitation source and vocal tract, which is digitally encoded and transmitted to the receiver. At the receiver, the decoded parameters are used to synthesize a signal which is perceptually similar to the original speech. The vocal excitation is either a broad-spectrum, quasi-periodic train of impulses
<table>
<thead>
<tr>
<th>CLASS</th>
<th>EXAMPLES</th>
<th>BIT RATE (kbps)</th>
<th>QUALITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis/Synthesis Coding</td>
<td>Channel Vocoder, Formant Vocoder, LPC Vocoder, Homomorphic Vocoder, Pattern Matching Vocoder, Phonetic Vocoder, Voice-Excited Channel Vocoder</td>
<td>0.8 - 2.4</td>
<td>Fair at the higher end of the bit rate, synthetic at the lower end of the bit rate</td>
</tr>
<tr>
<td>Waveform Coding</td>
<td>Time Domain:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pulse Code Modulation, Adaptive Differential PCM</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>PCM</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ADPCM</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adaptive Delta Modulation</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adaptive Predictive coding with noise shaping</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Frequency Domain:</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Subband coding</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adaptive Transform coding</td>
<td>9.6 - 16</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Phase Vocoder</td>
<td>7.2 - 9.6</td>
<td></td>
</tr>
<tr>
<td>Hybrid Coding</td>
<td>SBC or ATC with Harmonic scaling</td>
<td>8 - 9.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Voice-Excited LPC</td>
<td>4.8 - 9.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Residual-Excited LPC</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3.1 Summary of the bit rates and quality of the three categories of speech coding systems.

* (Commentary-grade waveform coders are for wideband speech. The rest are for narrowband speech.)
Figure 2.3.1 Generalized block diagram of a Vocoder
for voiced speech or a broad-spectrum, random noise signal for unvoiced speech. A voiced/unvoiced decision is therefore required to be extracted from the input signal and transmitted to the receiver. The vocal excitation information can be in the form of pitch period, fundamental or formant frequencies of the input signal depending on the model assumed. The information about the vocal tract varies from the short-term spectral envelope of the input signal to the coefficients of an all-pole filter. More detailed discussion of the various vocoding techniques are given in the subsequent sections.

Implicit in all vocoding procedures is the assumption that the source excitation and the vocal tract are linearly separable elements. Vocoder provide speech intelligibility rather than quality. Though they operate at the lower end of the transmission bit rate spectrum, they generally have the highest implementation complexity.

2.3.1.1 Channel vocoder (18-20)

The invention of channel vocoder by Homer Dudley in 1939 (18) initiated the analysis-synthesis technique to achieve bandwidth compression for speech coding. Figure 2.3.2 shows the system block diagram of a channel vocoder. The vocoder first divides the speech signals into frequency bands (or channels) by means of a bank of contiguous bandpass filters. The signals at the output of the filters are then rectified and lowpass filtered to produce signals which vary slowly
with time. The signals are perceptually significant as they represent the short-term amplitude spectrum of speech. They are sampled at Nyquist frequency, digitally encoded and transmitted to the receiver. Experiments showed that they need to be updated only once every 20 msec. The total number of channels is typically eleven with each channel having a bandwidth of 300 Hz. Each of the lowpass filters following the rectifiers has a bandwidth of 25 Hz. Therefore, the signals describing the short-term spectrum occupy a total bandwidth of less than 300 Hz. Substantial bandwidth compression is thus achieved.

The channel vocoder assumes that the voiced sounds are produced by exciting the vocal tract by a periodic pulse train and unvoiced sounds by random noise. Therefore voiced/unvoiced decision has to be extracted from the input signals. For voiced sounds, a frequency discriminator and a meter measure the fundamental frequency or pitch period, of the quasi-periodic signals. It represents the excitation information of the speech production model. The output signals of the filter-bank and the voiced/unvoiced decision are digitally encoded, multiplexed and transmitted to the receiver.

For the reconstruction of the speech spectrum at the receiver the excitation source, either from a pitch-modulated constant average power pulse generator, or from a broad-band noise generator, is applied to an identical set of bandpass filters. The input of these filters is amplitude modulated by the spectrum defining signals. The short-term spectrum approximating that measured at the transmitted is thus created.
Typical transmission bit rates of channel vocoders are between 2.4 and 4.8 kbps. The output speech exhibits perceptible degradations in speech naturalness and quality though intelligibility may be high. This is due to the following factors:

(1) Errors in voiced/unvoiced decisions

(2) Voiced sounds are synthesized from a pulse source whose waveform and phase spectrum do not reflect the actual details and changes of the real glottal waveform

(3) There is a lack of spectral resolution at the analysis stage due to the number, and thus the bandwidth of the analysing filters

(4) Noise source for unvoiced excitation is only at best a good approximation

2.3.1.2 Formant Vocoder (21,22)

The earliest parallel formant vocoder is illustrated in Figure 2.3.3. At the analyser, the input speech band is divided into four subbands. In each band, the average frequency of axis-crossing, F, and the average rectified and smoothed amplitude, A, corresponding to the formant frequency and amplitude, are measured at the receiver and transmitted to the synthesizer. The synthesizer contains three resonators which are controlled by F1, A1, F2, A2 and F3, A3.
Figure 2.3.2 System block diagram of channel vocoder (after Reference 19)

Figure 2.3.3 System block diagram of parallel-connected formant vocoder (after Reference 21)
Voiced excitation of the resonator is signalled by the voicing amplitude $A_v$. As in the channel vocoder, the frequency of the pulse source is described by $F_0$. Unvoiced excitation is determined by the amplitude $A_u$. The output of all the parallel filters are combined to yield the synthesized speech. Formant vocoder can also be implemented in cascade form using digital filters. Detailed discussions can be found in reference 22.

2.3.1.3 Linear Prediction (LPC) Vocoder

The linear prediction coding vocoder (Fig. 2.3.4) achieves speech production by exciting an all-pole recursive digital filter by either a periodic pulse train or random noise. It is a time-domain technique and therefore the difficult task of formant tracking as exists in the frequency-domain formant vocoder can be avoided. The adaptive $P$th order all-pole filter accounts for the vocal tract characteristics and its transfer function $H(z)$ is given by

$$H(z) = \frac{G_A}{\prod_{i=1}^{P} a_i z^{-i}} \quad (2.3.1)$$

where $P = 2L$ and $L$ specifies the number of formants needed to describe the short-term amplitude spectrum. $G_A$ determines the amplitude of the excitation. When a voiced or unvoiced sound is to be produced, the filter $H(z)$ is excited by a quasi-periodic or a random noise respectively.
Figure 2.3.4  System block diagram of LPC synthesizer
The coefficients $a_k$'s are determined by minimizing the mean square errors between samples of the input signals and the signal values estimated from a weighted linear sum of past values of the signal. That is, for every sample of the input speech $x_i$, an estimate $\hat{x_i}$ is formed such that

$$\\hat{x_i} = \sum_{k=1}^{P} a_k x_{i-k}$$

(2.3.2)

The coefficients, $a_k$'s, are determined by minimizing $E[(x_i - \hat{x}_i)^2]$ over an analysis interval of typically a few pitch period. They are given as the solution of the matrix equation

$$\phi A = \psi$$

(2.3.3)

where $A$ is the $P$-dimensional vector whose $k$th component is $a_k$, $\phi$ is a $(P \times P)$ covariance or autocorrelation matrix with $\phi_{ij}$ term given by

$$\phi_{ij} = \sum_{n=1}^{P} x_{n-i} \cdot x_{n-j}$$

(2.3.4)

or

$$\phi_{ij} = \phi_{|i-j|} = \sum_{n=1}^{P-|i-j|} x_n x_{n+|i-j|}$$

(2.3.5)

for $i,j = 1, \ldots, P$ and $\psi$ is a $P$-dimensional vector with the $j$th component $\psi_j = \phi_{jo}$. The vector $A$ determined by the covariance method does not ensure the stability of the inverse filter $H(z)$ because the diagonal elements of the covariance matrix are not necessarily identical and thus
the resulting poles of $H(z)$ may lie outside the unit circle $^{(25)}$. This problem can be avoided by using the autocorrelation method to calculate the prediction vector $A$. The advantage of the covariance method is that it provides a more accurate estimation of the predictor coefficients because more samples are used for the formation of the covariance function than that in the autocorrelation method $^{(25)}$. An alternative implementation of linear prediction of speech is by means of an adaptive lattice predictor proposed by Itakura and Saito $^{(26)}$. As the partial correlation (PARCOR) coefficients are ensured to have magnitude less than one, the stability of the inverse filter is guaranteed.

One problem with LPC vocoder is the extraction of the pitch period for voiced speech. Though there are numerous pitch extraction algorithms $^{(28)}$ existing in the literature, none of them is known to be perfect especially in a noisy environment. In ideal laboratory conditions, reasonable quality speech can be obtained using the LPC vocoding technique at the bit rate of 4.8 kbps and below.

2.3.2 Time-Domain Coders

2.3.2.1 Pulse Code Modulation (PCM) $^{(4)}$

The very basic technique for converting analogue signals into digital form is the technique of Pulse Code Modulation (PCM) invented by Reeves $^{(3)}$ in 1938. It is still the most widely used technique for digitizing analogue speech signals for telephony application. When it
is used for feeding analogue signals into computers or other digital equipment for subsequent processing it is known as analogue-to-digital conversion (ADC). Detailed analysis of PCM can be found in reference (4) by Cattermole.

The input signal $x(t)$ is first bandlimited to $f_{\text{max}}$ by means of a low-pass filter, and sampled at a frequency $f_s$ equal to or greater than the Nyquist frequency $2f_{\text{max}}$. The samples are quantized in amplitude by rounding off each sample value to one of a set of several discrete values. In a B-bit quantizer, the number of discrete amplitude level is $2^B$. These discrete amplitude levels are represented by distinct binary words of length $B$. For decoding, the binary words are mapped back into an amplitude-time pulse sequence which is lowpass filtered to the cut-off frequency $f_{\text{max}}$ to reproduce the recovered analogue speech $x(t)$.

(i) **Uniform Quantization**

The most important element in PCM is perhaps the quantizer. Let the step-size of a zero-memory and time-invariant uniform quantizer be denoted by $\Delta = x_i - x_{i-1}$ where $x_i$'s are the input decision thresholds. If the quantizer is designed to match the input signal power, so that no overloading of the quantizer can occur, and if the number of quantization level is large, the quantizer error $E$ can be described by an uniform distribution given by

$$p(E) = \frac{1}{\Delta} \quad \frac{\Delta}{2} \leq E \leq \frac{\Delta}{2}$$

(2.3.6)
The mean-square value of the quantization error is:

\[ N_q^2 = \int_{-\Delta/2}^{\Delta/2} E^2 p(E) dE \]

\[ = \frac{\Delta^2}{12} \quad (2.3.7) \]

For an input signal with rms value of \( \sigma_x \), the signal-to-noise ratio (SNR) in dB is then defined by

\[ \text{SNR} = 10 \log_{10} \left( \frac{\sigma_x^2}{2N_q} \right) \]

\[ = 10 \log_{10} \left( \frac{\sigma_x^2}{(\Delta^2/12)} \right) \quad (2.3.8) \]

If the input samples have a zero mean Gaussian or Laplacian pdf with standard deviation \( \sigma_x \), only 0.01% or 0.35% of the samples fall outside the \( \pm 4 \sigma_x \) range. The step-size of the uniform B-bit quantizer is simply given by

\[ \Delta = \frac{8 \sigma_x}{2^B} \quad (2.3.9) \]

From equations (2.3.8) and (2.3.9)

\[ \text{SNR} = 6B - 7.2 \text{ (dB)} \quad (2.3.10) \]
Equation (2.3.10) is only true for fine quantization, i.e. for large value of $B (x \neq 0)^{(4)}$. It shows that SNR value increases linearly with $B$. Under fine quantization condition, quantization noise has a broad frequency spectrum and it is uncorrelated with the input signal. Clearly, a uniform quantizer does not achieve the best possible SNR performance as the amplitude distribution of the input samples, which is not uniform in most cases of interests, is not made use of in its design. More efficient quantizers have their input-output characteristics matched to the amplitude pdf of the input samples or employs some form of adaptation strategy to maximize their SNR performance.

(ii) **Non-Uniform Quantization**$(4,29)$

Non-uniform quantization is characterised by fine quantization steps for the very frequently occurring low amplitude samples in speech and coarse quantizing steps for the occasional samples with large amplitude excursions (Figure 2.3.5). Average distributions of speech amplitudes are decreasing functions of amplitude and non-uniform quantization constitutes a direct utilization of this property. Two non-uniform quantizers widely used in commercial telephony applications are A law and $\mu$ law PCM$^{(4)}$. They are functionally equivalent to uniform quantization of a logarithmically compressed input.

For an input signal $x$, both laws are symmetrical about $x = 0$. For $x > 0$ and $x_{\text{max}} = V$, the compressed signal $x_c$ are defined as follows$^{(4)}$:
\[ \mu\text{-law : } x_c = \frac{V \ln (1 + \mu x/v)}{\ln (1 + \mu)}, \quad 0 < x < v \quad (2.3.11) \]

\[ \text{A-law : } x_c = \frac{Ax}{1 + \ln A}, \quad 0 < x < V/A \]

\[ = \frac{V (1 + \ln(Ax/V))}{1 + \ln A}, \quad V/A < x < V \quad (2.3.12) \]

A commonly used value for \( \mu \) is 255 and \( A \) takes a value of 86 for a 7-bit PCM speech coder. The use of a logarithmic characteristic allows the quantizer to span the large dynamic range of typical speech signals. An 8-bit log-quantizer was found to have the same dynamic range as a 12-bit uniform quantizer\(^{(4)}\).

The dynamic range of speech signals encountered in typical voice communication systems can be as much as 40 dB. Though the existing networks make use of time-invariant log PCM as a solution, considerable improvement can be obtained by designing a quantizer with time-varying step-size to adapt to the non-stationary properties of speech signals. Quantizers with time-varying step-sizes are known as adaptive quantizers (AQ). The adaptation strategy can be either forward (AQF) or backward (AQB). In AQF, side-information related to the step-size of the quantizer has to be transmitted, normally at the time interval of about 16-20 msec, to the receiver. In AQB, the adaptation is based on the previously decoded sample(s) and therefore, no side-information is required.
(iii) Adaptive Quantization

(a) Adaptive Quantization Forward (AQF)

In forward adaptive quantization\(^{(30)}\), the step-size \( \Delta \) of the quantizer is calculated over a block of \( N \) samples by the equation

\[
\Delta = \alpha \sqrt{\frac{\sum_{i=1}^{N} x^2(i)}{N}}
\]  

where \( x(i) \) is the \( i \)th-sample of the input \( x \) in time and \( \alpha \) is a constant which depends on the number of levels of the quantizer. The quantizer step-size is then maintained for the quantization of the \( N \) samples. It is updated again according to equation (2.3.13) for the next block of \( N \) samples. Clearly, the step-size has to be transmitted to the receiver as side-information. Additional bit rate is therefore incurred. Another disadvantage of AQF is the \( N \)-sample delay introduced to the coder. However, the adaptation procedure ensures that the quantizer step-size is always matched to the input power and thus substantial improvement over time-invariant quantizer can be achieved.

(b) Adaptive Quantization Backward (AQB)\(^{(31-33)}\)

The most well known quantizer which uses backward adaptation strategy is the one-word-memory adaptive Jayant's quantizer\(^{(31)}\) (AQJ). The step-size of AQJ shrinks or expands at every sampling instant like an accordion. Figure (2.3.6) shows the input-output characteristics of
Figure 2.3.5 The input-output characteristics of a non-uniform quantizer

Figure 2.3.6 Adaptive quantizer characteristics for
(a) Low signal power and
(b) Large signal power
AQJ at 2 time-instants. The adaptation rule can be simply described as below.

Consider at the ith sampling instant, the step-size of a B bit uniform quantizer to be \( \Delta_i \) and its output level \( x_i \), i.e.

\[
x_i = \pm p_i \frac{\Delta_i}{2} \quad p_i = 1, 3, 5, \ldots, 2^{B-1}
\]

(2.3.14)

The quantizer step-size \( \Delta_{i+1} \) at the next sampling instant is obtained by multiplying \( \Delta_i \) by a fixed expansion-compression coefficient \( M_\Delta \) which is determined from the previous quantizer output level, i.e.

\[
\Delta_{i+1} = \Delta_i \cdot M_\Delta[p_k]
\]

(2.3.15)

where \( M_\Delta[p_k] \) is a function of \( p_k \).

\( M_\Delta \) is one of the \( 2^{B-1} \) fixed coefficients. It has a value of less than but close to unity when \( p_k \) corresponds to one of the inner quantization levels. For outer quantization levels, the value of \( M_\Delta \) lies between 1 and 2.5. With this strategy, the rate of increase in step-size is greater than the rate of decrease and thus the occurrence of possible overload errors is minimized. Since the step-size adaptation follows the quantizer output rather than input, no step-size information has to be explicitly communicated to the receiver. More detailed discussions of AQJ are presented in Chapter 3.
2.3.2.2 Vector Pulse Code Modulation (VPCM) (34,35)

Conventional PCM digitize speech waveform sample by sample with uniform or logarithmic spacing of quantization levels. A relatively new direction in source coding is the vector quantization of a block of samples of a waveform. In vector PCM (VPCM), a block of consecutive samples forms a vector that is treated as one entity. The vector is encoded with a binary word, and an approximation of the original vector is generated using only this binary word at the receiver. Vector quantization is essentially a pattern matching technique. Each input vector is encoded by comparisons with a set of codevectors stored in a 'codebook'. It is represented by one of the codevectors which is identified as the most 'similar' to it according to a suitable fidelity measure. The codebook size (number of patterns) is a critical parameter which determines the encoding complexity needed for searching through the codebook, the memory required to store the codebook in both transmitter and receiver, and the bit rate for digital transmission.

Let \( N \) be the number of patterns in the codebook, \( k \) the dimension of the vector, and \( r \) the transmission rate in bits/sample. Clearly \( r \) is given by:

\[
    r = \frac{\log_2 N}{k} \tag{2.3.16}
\]

When the number of patterns, \( N \), is sufficiently large the signal-to-noise ratio for VPCM can be shown to be given by (35):
\[ \text{SNR} = 6 \log_2(N)/k + C_k \]  

(2.3.17)

where \( C_k \) is a constant depending on the dimension \( k \), and \( \text{SNR} \) and \( C_k \) are expressed in dB. The SNR of VPCM therefore increases at the rate of \( \frac{6}{k} \) dB for each doubling of the codebook size. Equation (2.3.17) reduces to Equation (2.3.10) for the case of scalar PCM with \( k = 1 \) and \( C_k = -7.2 \). For values of \( k \) greater than 1, the correlation between the components of the vector is exploited and this leads to an increase in the value of \( C_k \). Experiments with speech signals show that \( C_2 \) is greater than \( C_1 \) by more than 3 dB and \( C_8 \) is greater than \( C_1 \) by more than 7 dB \((35)\). VPCM is clearly an attractive improvement over standard PCM but it has a much higher coder complexity.

2.3.2.3 Differential Pulse Code Modulation \((30, 36-38)\)

Speech signals sampled at the Nyquist rate exhibits very high correlation between successive samples. This is due to the lowpass characteristics of the speech spectrum especially during voiced speech. In differential coding, efficiency is obtained by quantizing the sample-to-sample difference of the speech signal rather than the signal itself. As the sample correlation is high, the rms level of the difference signal is much less than that of the signal. Thus, a smaller quantizer step-size may be used which leads to smaller overall quantization noise. Differential coding was first proposed by C. C. Cutter in 1952 \((36)\). Since then, it is universally known as Differential Pulse Code Modulation or DPCM in short.
The schematic diagram of DPCM is illustrated in Figure 2.3.7. The bandlimited analogue signal $x(t)$ is sampled at the Nyquist rate to produce a sequence of samples $\{x_i\}, \ i = 1, 2, \ldots, \infty$. At the same time, based on the previously decoded samples $\{\hat{x}_i\}$, a linear predictor forms $\{y_i\}$ as an estimate of $\{x_i\}$. The difference sequence is then quantized, binary encoded and transmitted to the receiver. From the diagram, the following equations can be easily obtained:

\[
\begin{align*}
    e_i &= x_i - y_i \\
    \hat{e}_i &= e_i + q_i \\
    \hat{x}_i &= \hat{e}_i + y_i \\
    \hat{x}_i &= x_i + q_i
\end{align*}
\]

where $q_i$ is the quantization noise introduced by the quantizer. Notice that the locally decoded sample $\hat{x}_i$ is the sum of the input sample $x_i$ and the quantization noise $q_i$.

The predictor in the feedback loop of a DPCM system usually takes the form of an $N$th-order linear transversal filter with coefficients $a_i$'s ($i = 1, \ldots, N$). The output of the filter or predictor, $y_i$, can be expressed as

\[
y_i = \sum_{k=1}^{N} a_k \hat{x}_{i-k}
\]

For a first-order predictor, $a_1$ is set to $C_1$ which is the first-shift autocorrelation coefficient of the input signal. For speech, it has
\[ \hat{e}_i = e_i \cdot q_i \]

Figure 2.3.7 System block diagram of the DPCM codec
a value of greater than 0.8, but less than 1.0 of course, and it is usually called leaky integrator. For narrowband speech samples at 8 kHz, the value of $C_k$ is around 0.85.

(i) **Adaptive DPCM (ADPCM)**

Recall that the aim of a differential coder is to minimize the quantization noise of the decoded speech by exploiting the sample-to-sample correlation inherent in speech signals. To minimize $q_i$, the step-size of the quantizer has to adapt to the changing rms level of the input signal $e_i$ and adaptive quantizers like AQF and AQJ may be employed. As lower rms level of $e_i$ leads to lower rms level of $q_i$, further improvement in DPCM SNR performance can be achieved by employing an adaptive predictor which gives the closest possible predicted value $y_i$ of $x_i$ so that $e_i$ is minimized. A leaky integrator or a fixed Nth-order predictor obviously does not provide the best solution as speech signal is a non-stationary signal in the long-term.

In the frequency domain, the non-stationary properties of speech are equivalent to the dynamic characteristics of the spectral envelope. However, within a short time frame of about 20 msec, speech can be considered statistically stationary with its amplitude spectral envelope defined by the frequency characteristics of the vocal tract. Predictors able to tract the time-varying spectral characteristics of speech are termed adaptive predictors. As they have to be implemented at
both the transmitter and receiver, their adaptation is usually derived from the locally decoded sequence \( \{\hat{x}_i\} \) either sequentially or on a block-by-block basis. Alternatively, if the adaptation is derived from the input samples, side-information regarding the predictor coefficients has to be transmitted to the receiver.

(ii) Adaptive Prediction \(^{37,39-46}\)

The objective of the predictor in an ADPCM speech coder is to provide as accurate prediction of the future input speech samples as possible so that the power of the error sequence obtained by subtracting the predicted sequence from the input sequence is minimum. In the case of fixed predictor, the prediction coefficients are designed on the basis of the long-term signal statistics. Obviously, this type of prediction is not the most efficient because speech signal is non-stationary in the long-term. There are many ways the prediction coefficients can be made to adapt to the input characteristics. In forward block adaptive prediction \(^{30}\), the coefficients are derived directly from a block of buffered input samples. As the receiver has no knowledge of the input sample values, the prediction coefficients have to be transmitted as side-information. This leads to an increase in transmission bit rate and therefore the adaptation can only be carried out over a block of samples of about 20 msec. In the category of backward sequential adaptive prediction \(^{39-45}\), the adaptation of the prediction coefficients is based on the locally decoded samples available
both at the encoder and decoder. As such, no transmission of side-information pertaining to the adaptation of the prediction coefficients is required. Due to this advantage, the adaptation process can be carried out sequentially, i.e. on a sample-by-sample basis. The generalized block diagrams of the ADPCM systems employing the forward and backward prediction adaptation strategies are shown in Fig. 2.3.8a and b respectively.

(a) *Forward Block Adaptive Prediction (FBAP)*$^{(37)}$

In this method, the prediction coefficients are computed from the buffered $W$ samples of the original speech input. If $N$ is the order of the predictor, its coefficients are designed to minimize the block error energy, $\sigma^2$, given by

$$\sigma^2 = \sum_{i=1}^{W} (x_i - \sum_{k=1}^{N} a_k x_{i-k})^2$$

(2.3.23)

where $x_i$'s and $a_k$'s are input sample values and prediction coefficients respectively. To obtain the values of $a_k$'s, the derivative of $\sigma^2$ with respect to $a_k$'s is set to zero to yield the normal equations

$$\sum_{k=1}^{N} a_k (\sum_{i=1}^{W} x_{i-k} x_{i-r}) = \sum_{i=1}^{W} x_{i-r} x_i$$

(2.3.24)

where $r = 1, 2, \ldots, N$. 
Figure 2.3.8a Generalized block diagram of an ADPCM system using Forward Block Adaptive Prediction (FBAP)

Figure 2.3.8b Generalised block diagram of an ADPCM system using Backward Sequential Adaptive Prediction (BSAP)
The solution of Equation (2.3.24) depends on how $x_i$'s are specified.

In the autocorrelation method, $x_i$'s are defined by

$$x_i = \begin{cases} x_i & 1 \leq i \leq W \\ 0 & \text{otherwise} \end{cases}$$  \hspace{1cm} (2.3.25)$$

In the covariance method, the values of $x_i$'s outside the window length $W$ are not set to zeros.

In general, the autocorrelation method is preferred as its solution leads to a guaranteed stable inverse filter at the receiver (25).

Furthermore, the normal equations can be solved by an elegant recursive algorithm derived by Durbin (24). In solving for the coefficients of an Nth order predictor, Durbin's algorithm also computes the solutions for all predictors of order less than $N$. As for its processing complexity, only $N^2$ operations and $2N$ storage locations are required.

(b) **Backward Sequential Adaptive Predictor (BSAP)** (39-44)

In the general techniques of Backward Sequential Adaptive Prediction, the adaptation of the $k$th prediction coefficient at the $(i+1)$th sampling instant can be described by the equation

$$a_{i+1,k} = a_{i,k} - g \frac{\partial \text{EF}}{\partial a_{i,k}}$$ \hspace{1cm} (2.3.26)$$

where EF refers to the error function to be minimized and $g$ controls the rate of adaptation of the algorithm. Predictors employing this
kind of adaptation algorithm are sometimes referred to as gradient predictors as the second term of Equation (2.3.26) is the gradient of the error function with respect to $a_{i,k}$. If the error function to be minimized is the mean square error, i.e. $EF \triangleq \langle \hat{e}_i^2 \rangle$, the kth component of the gradient is given by

$$\frac{\partial (EF)}{\partial a_{i,k}} = \frac{\partial}{\partial a_{i,k}} \langle \hat{x}_i - \sum_{k=1}^{N} a_{i,k} \hat{x}_{i-k} \rangle^2 \quad (2.3.27)$$

or

$$V(EF) = -2 \begin{bmatrix} \hat{x}_{i-1} \\ \vdots \\ \hat{x}_{i-N} \end{bmatrix} \begin{bmatrix} \hat{x}_{i-2} \\ \vdots \\ \hat{x}_{i-N} \end{bmatrix} = -2 \begin{bmatrix} \hat{x}_i - \sum_{k=1}^{N} a_{i,k} \hat{x}_{i-k} \\ \vdots \\ \hat{x}_{i-N} \end{bmatrix} \begin{bmatrix} \hat{x}_i - \sum_{k=1}^{N} a_{i,k} \hat{x}_{i-k} \\ \vdots \\ \hat{x}_{i-N} \end{bmatrix} \quad (2.3.28)$$

where $\hat{x}_i$'s are the locally decoded samples. From Equations (2.3.26) and (2.3.27) we have

$$a_{i+1,k} = a_{i,k} + h \hat{x}_{i-k} \hat{e}_i \quad (2.3.29)$$

where $h = 2g$. The value of $h$ is optimized using computer simulation.

(b.1) **Stochastic Approximation Prediction (SAP)**

The Stochastic Approximation Prediction is a Backward Sequential Adaptive Prediction technique proposed by Cummiskey (45). It is also based on the Equation (2.3.29). In this algorithm, instead of keeping
h as a fixed constant, it is made to vary with \( \hat{x}_i \)'s by

\[
h = \frac{A}{B + \zeta(\hat{x}_i, M)}
\]  

(2.3.30)

where \( A, B \) are constants and \( \zeta(\hat{x}_i, M) \) is a function of the \( M \) previous locally decoded samples. It is defined by

\[
\zeta(\hat{x}_i, M) = \frac{1}{M} \sum_{k=1}^{M} \hat{x}_{i-k}^2
\]  

(2.3.31)

The term \( B + \zeta(\hat{x}_i, M) \) is a form of automatic gain control that equalizes the adaptation rate of the algorithm as the input speech power level varies. As it increases with an increase in power, \( h \) decreases and overcorrections of the \( a_{i+1,k} \) coefficients are avoided to prevent the occurrence of large prediction error. For silence or unvoiced segments of speech input, \( \zeta(\hat{x}_i, M) \ll B \) and Equation (2.3.30) maintains a finite value. Thus the bias term \( B \) compensates for the low level input signals. Henceforth, \( h \) will be replaced by \( P_i(\hat{x}) \) since it is variable and changes at every sampling instant, i.e.

\[
P_i(\hat{x}) = \frac{A}{B + \frac{1}{M} \sum_{j=i-M-1}^{i-1} \hat{x}_j^2}
\]  

(2.3.32)

(b.2) **Sequential Gradient Estimation Predictor (SGEP)**\(^{(37,40,41)}\)

Another example of Backward Adaptive Prediction is the SGEP algorithm proposed by Evci and Xydeas. It also updates the prediction coefficients
according to the general formula given by Equation (2.3.29), but the gradient term is replaced by the term \( P_i(\hat{x})d_{i,k}/k^\alpha \), and the kth prediction coefficient at the \((i+1)\)th sampling instant is now obtained as

\[
\hat{d}_{i+1,k} = \hat{d}_{i,k} - P_i(\hat{x}) \frac{d_{i,k}}{k^\alpha}, \quad k = 1, 2, \ldots, N, \tag{2.3.33}
\]

or in the vector form,

\[
A_{i+1} = A_i - P_i(\hat{x})K D_i \tag{2.3.34}
\]

where \( K \) is an \((N \times N)\) diagonal matrix given by

\[
K = \begin{bmatrix}
1 & 0 & \cdots & 0 \\
0 & 2^{-\alpha} & 0 & \cdots \\
\vdots & \ddots & \ddots & \ddots \\
0 & \cdots & \cdots & N^{-\alpha}
\end{bmatrix} \tag{2.3.35}
\]

The term \( d_{i,k} \) will be explained later.

\( P_i(\hat{x}) \) is given by Equation (2.3.32) and the term \( k^{-\alpha} \), with \( 0 < \alpha < 1 \), serves as a weighting factor so that the higher order prediction coefficients receive smaller modification than the lower order coefficients. This is in agreement with the observation that the lower order coefficients are more important in determining the performance of an ADPCM system and
therefore they should receive higher adaptation than the higher order coefficients.

The most important factor in controlling the performance of SGEP is the term \( d_{i,k} \). The value of \( d_{i,k} \) used in updating the kth coefficient, at the ith sampling instant, is determined by the two prediction error functions \( EF_{i,2k-1} \) and \( EF_{i,2k} \). For each coefficient, \( d_{i,k} \) is given by

\[
d_{i,k} = EF_{i,2k-1} - EF_{i,2k}
\]

(2.3.36)

while the two error functions, \( EF \), are formed as follows:

At the sampling instant \( i \), consider the coefficient \( a_{i,1} \). It is increased by a positive number \( S_{i,1} \) which is given by

\[
S_{i,k} = \frac{1}{Gk^\beta}
\]

(2.3.37)

where \( G \) and \( \beta \) are constants, and \( G > 1 \), \( 0 < \beta < 1 \). Having the value of \( a_{i,1} \) modified by \( S_{i,1} \), a predicted value \( y_{i,1} \) of the input sample \( x_1 \) is obtained. \( a_{i,1} \) is then decreased by the same amount, \( s_{i,1} \) and another predicted output \( y_{i,2} \) is obtained. In the same way, \( a_{i,2}, \ldots, a_{i,N} \) are modified by \( S_{i,2}, \ldots, S_{i,N} \) to form \( y_{i,3}, y_{i,4}, \ldots y_{i,2N-1}, y_{i,2N} \). When absolute error criterion is used, the error functions are then given by
\[
\begin{align*}
\text{EF}_{i,1} &= |x_i - y_{i,1}| \\
\text{EF}_{i,2} &= |x_i - y_{i,2}| \\
&\quad \vdots \\
\text{EF}_{i,2N-1} &= |x_i - y_{i,2N-1}| \\
\text{EF}_{i,2N} &= |x_i - y_{i,2N}| \\
\end{align*}
\]

(2.3.38)

Alternatively, the mean square error criterion can be employed. Once \(d_{i,k}\) is determined, the prediction coefficients can be updated using Equation (2.3.33).

Other backward sequential predictors with the general form of Equation (2.3.29) have been studied by Gibson\(^{(39, 42-44, 47)}\), Moye\(^{(48)}\), Jones, Cohn and Melsa\(^{(39, 49)}\), Qureshi and Forney\(^{(50)}\). One disadvantage associated with backward sequential predictors in ADPCM is the risk of instability of the inverse filter when transmission errors occur. As the adaptation is based on the locally decoded samples, transmission errors will cause a mismatch between the adaptive predictor at the receiver and transmitter. Theoretically, the effect can perpetuate for ever unless some form of leakage is introduced to stop its propagation. However, the introduction of predictor leakage to dissipate the effect of transmission errors means that the adaptation rate of the predictor is slowed down and thus leads to a decrease in system performance.
All the backward sequential predictors described so far have the conventional linear transversal filter structure. A different way to implement it is by means of a lattice structure (51-54). An inverse filter with a lattice structure has the advantage that stability is guaranteed if the PARCOR or reflection coefficients are constrained to be of magnitude less than one (51). To dissipate the effect of transmission errors, a leakage factor is still required in the adaptation algorithm. More details of lattice predictors are given in Chapter 3.

2.3.2.4 Delta Modulation (DM)

DPCM coders remove the redundancy between the successive samples by means of various forms of sophisticated predictors. Together with adaptive quantization strategies, they achieve higher SNR performance than those with fixed prediction and quantization. The relative complexity of an adaptive predictor in a differential coder can be avoided by increasing the sample-to-sample correlation through over-sampling, i.e. sampling frequency > 4 times the Nyquist frequency so that a simple single or double integrator is sufficiently good as a replacement. Differential coders employing over-sampling, simple prediction and two level quantization strategies are termed Delta Modulation or Delta Modulators (DM). A detailed presentation of DM system can be found in Reference 55.
The schematic diagram of a Linear Delta Modulator (LDM), which is the simplest form of Delta Modulator, is illustrated in Figure 2.3.9. The input signal $x(t)$, bandlimited to $f_c$, is sampled at a frequency $f_p$ which is much higher than the Nyquist frequency. An error sequence $\{e_i\}$ is obtained by subtracting the predictor sequence $\{y_i\}$ from the input sequence $\{x_i\}$. $e_i$ is then quantized by a two level quantizer with a step size $\delta$. The local decoder forms $y_i$, the prediction of $x_i$, by integrating the output of the quantizer, i.e.:

$$y_i = y_{i-1} + a \delta \text{sign}(e_i)$$

(2.3.39)

where \text{sign}(\cdot) means the sign of $\cdot$ and $a = 1$ for an ideal integrator and $a < 1$ for a leaky one. The output of the quantizer $\pm \delta$ is transmitted as a one-bit word. The decoder at the receiver is identical to the local decoder at the transmitter. The recovered $\hat{x}(t)$ is obtained by passing $\{Y_i\}$ through a lowpass filter having a cut-off frequency $f_c$ to remove the out-of-band noise.

Two types of distortion, namely granular noise and slope overload, can occur in a Delta Modulator. The former is determined by the step-size of the quantizer $\delta$. The latter is caused by the inability of the encoder to follow the signal when its slope magnitude $|\dot{x}|$, exceeds the ratio of step-size to sampling period, i.e.

$$|\dot{x}| > \delta/T$$

(2.3.40)

These two types of distortion are illustrated in Figure (2.3.10).
Figure 2.3.9  The Linear Delta Modulator
(a) Encoder
(b) Decoder

Figure 2.3.10  Waveforms for a Delta Modulator with
single integration
(after Reference 57)
Granular noise can be made small by using a small step-size and slope overload can be reduced by a large step-size or by increasing the sampling frequency. The latter however, increases the transmission bit rate. For a given bit rate, the step-size is selected to give a compromise between granular distortion and slope overload.

(i) **Adaptive Delta Modulation (ADM)**\(^{(58-64)}\)

To minimize the granular noise and slope overload distortions without increasing the transmission bit rate, the step-size has to be varied according to a prescribed logic. The step-size control logic may be discrete or continuous and it may act with a short-time constant, i.e. sample-by-sample, or with a time constant of syllabic duration. Normally, it is controlled by information contained in the transmitted bit stream if no transmission of side-information is required.

(a) **Constant Factor Delta Modulation (CFDM)**

The simplest form of adaptive Delta Modulation is perhaps the Constant Factor Delta Modulation (CFDM) proposed by Jayant\(^{(59)}\). The coder responds to the instantaneous variations in the analogue signal and is suitable for encoding both speech and television signals. Figure 2.3.11 shows the schematic block diagram of the coder. In its operation, successive bits \(b(n)\) and \(b(n-1)\) are compared in order to detect probable slope overload (\(b(n) = b(n-1)\)) or probable granularity (\(b(n) \neq b(n-1)\)).
The step-size \( A \) adapts according to

\[
A(n) = A(n-1) M^{b(n)b(n-1)} \quad M \geq 1
\]  

(2.3.41)

i.e. \( A \) is either multiplied by \( M \) or \( \frac{1}{M} \) at each instant. The rate of step-size adaptation is governed by the single factor \( M \). For \( m = 1 \), the coder is equivalent to LDM. Jayant obtained an optimum value of \( M = 1.5 \) for toll quality speech at 60 kbps. At the bit rate of 32 kbps, degradation in quality is perceptible and a more complicated form of step-size adaptation is required. To improve the performance of CFDM, Steel and Kyaw\(^{(60)}\) employ the present bit, \( b(n) \), and the immediate past two bits, \( b(n-1) \) and \( b(n-2) \), for the control of the step-size. The coder is termed second order CFDM (SCFDM). The three bits in SCFDM constitute eight possible patterns which are grouped in complementary patterns to give four different multipliers. For Gaussian input signal bandlimited to 3.1 kHz, SCFDM achieved a 4.5 dB improvement in SNR measurement over CFDM.

(b) **Continuously Variable Slope Delta Modulation (CVSD)**\(^{(61)}\)

Continuously Variable Slope Delta Modulation, also known as syllabically companded Delta Modulation, employs a step-size adaptation strategy which depends on the syllabic rate of the speech waveform. The definition of syllabic rate in this context defers from its more well-known definition found in phonology where a syllable has a typical duration of about 100 msec. It corresponds to the pitch rate in the time waveform of speech of about 5-10 msec.
The schematic block diagram of CVSD is illustrated in Figure 2.3.12.

The operation of the coder can be described by the following equations:

\[
\hat{x}_{i+1} = \alpha \hat{x}_i + (1 - \alpha) \Delta_i e_i, \quad (2.3.42)
\]

\[
e_i = \text{sign of } (x_i - \hat{x}_i), \quad (2.3.43)
\]

\[
\Delta_{i+1} = \beta \Delta_i + (1 - \beta)(V + V_1), \quad (2.3.44)
\]

where \( \hat{x}_i \) is the estimate of the incoming signal,
\(
\alpha \) is the leakage factor associated with the estimate integrator,
\(
\beta \) is the leakage factor associated with the step-size integrator,
\( \Delta_i \) is the \( i \)th instant step-size,
\( V \) is the positive constant when 3 consecutive outputs from the CVSD encoder are identical and 0 otherwise,
and \( V_1 \) is another positive constant added to \( V \) to ensure that the minimum step-size is non-zero.

When optimized with respect to narrowband speech input, the time constants \( \tau_1 \) of the step-size integrator and \( \tau_2 \) of the estimate integrator in CVSD were found to be \( \tau_1 = 5.69 \text{ msec} \) and \( \tau_2 = 1 \text{ msec} \), which give

\[
\beta = \exp \left(1 - \frac{1}{F_s}\right)/(5.69 \cdot 10^{-3}) \quad (2.3.45)
\]

\[
\alpha = \exp \left(1 - \frac{1}{F_s}\right) / (10^{-3}) \quad (2.3.46)
\]
Figure 2.3.11 Jayant's one-word memory Constant Factor Delta Modulator (CFDM)

Figure 2.3.12 Block diagram of the continuously variable Slope Delta Modulator (CVSD)
When $F_s = 16$ kHz, $\beta = 0.99$ and $\alpha = 0.94$. The use of syllabic adaptation enables the Delta Modulator to tract the speech envelope. The use of the leaky step-size integrator also improves the robustness of the coder to transmission errors. When compared with CFDM under noisy channel conditions, CVSD was found to be less sensitive to transmission error\(^{61}\).

Other variations of ADM can be found in references 62 to 64. A theoretical study of the SNR performance of linear DM with Gaussian signals as input is given by Steele in reference 65. The stability issue of a double integration delta modulation is examined by Nielson in Reference 66.

2.3.3 Frequency-Domain Coders

In time-domain waveform differential coding techniques, redundancies in speech are removed by means of a fixed or adaptive predictor followed by quantization of an error sequence which is obtained after subtracting a prediction sequence from the input sequence samples. Time-domain coders are characterised by the quantization of full-band signals where each sample receives the same number of bits for quantization. The complexity of this type of coder is low to moderate and they are usually amenable to implementation using DSP devices.

In the second category of waveform coding techniques, i.e. frequency-domain coders, the encoder sub-divides the full-band speech signal
into a set of separate frequency-domain components followed by the encoding of these components according to perceptual criteria\(^{(67)}\). These systems have the great advantage that the number of bits can be distributed dynamically in such a way that the perceptually more important signal components received more accurate quantization than those which do not contribute much to the subjective quality. Recent research\(^{(67-72)}\) have shown that this second approach provides better quality decoded speech than most of the time-domain coders. However, the improvements are achieved through greater system complexity and delays due to the analysis and synthesis procedure required of the coders and the need to employ adaptive bit allocation algorithms.

The two most important frequency-domain coding techniques are Subband Coding (SBC)\(^{(68-70, 74-83)}\) and adaptive Transform Coding (ATC)\(^{(70-72)}\). In SBC, the sub-division of speech spectrum is accomplished through the use of filter-bank. In the more complex system of ATC, the function of filter-bank is effectively realized by means of a discrete transformation operating on large successive blocks of input data. Subband Coding is a broadband analysis-synthesis system which divides the full-band speech signal into only 4 to 16 subbands. The generally large-blocksize (\(\geq 128\) samples) transformation applied in ATC produces frequency components with much finer frequency resolution and it is therefore a narrowband analysis-synthesis technique.

2.3.3.1 Subband Coding (SBC)

In Subband Coding, the full-band speech signal is sub-divided into
typically 4 to 16 subbands by a bank of bandpass filters (Figure 2.3.13). The first step to process the subband signals is to translate them to the baseband by a modulation process equivalent to single-sideband translation. They are then down-sampled to their Nyquist rate and digitally encoded using ADPCM or APCM coders. The process of dividing the full band signal enables the bit allocation and coding of subband signals according to perceptual criteria. At the receiver, the subband signals are decoded and modulated back to their original frequency positions. After passing through the synthesis filter-bank, they are summed to give a replica of the original speech signal. The design concept of Subband Coding offers the following important advantages which do not exist in time-domain waveform coders:

1. The quantization noise is confined to each band and thus masking of one frequency range of the speech signal by quantization noise in another frequency range can be prevented;

2. Separate adaptive encoding can be employed in each band by exploiting the different properties of subband signals;

3. By appropriately allocating the bits in different bands, the spectral shape of the overall output noise can be controlled to give the best possible perceptual effect.

In general, more bits are allocated to the lower frequency bands
where pitch and formants must be accurately preserved. However, as
the speech signals in the higher frequency bands are fricative and
noise-like, fewer bits are found to be perceptually sufficient.

The earlier design problem with Subband Coding was the complexity
of the bandpass filters. Very high order (≥ 512) FIR filters
were used to achieve steep roll-offs at the transition regions so
that frequency aliasing of the subband signals, due to the down sampling
process, was not too high and perceptually annoying. Large order FIR
filters imply large processing requirement and longer coder delay.
The invention of Quadrature Mirror Filter (QMF) by Estaban alleviates the stringent requirement of the filters. With QMF, aliasing-
free reconstruction at the receiver was theoretically possible with
low-order (below 32) FIR filters. Since then, Subband Coding has
received great attention for the coding of narrowband speech at the bit
rate of 16 kbps and below.

In one of the earliest Subband Coding investigations, a 4-band subband
coder offers almost 100% preference in subjective listening tests over
a conventional ADPCM coder at the bit rates below 22 kbps. Similarly, at
the bit rates of 9.6 kbps and below, Subband Coding was found to have
a clear perceptual advantage over ADM operating at the same bit rate.

2.3.3.2 Adaptive Transform Coding (ATC)

In Transform Coding (TC), the input speech signal is first segmented
into successive blocks of N samples each. The N samples \( x_i \), \( i = 1, \ldots, N \), are then linearly transformed by an \((N \times N)\) orthogonal matrix
to yield the N transform coefficients \( y_i \), \( i = 1, \ldots, N \), i.e.

\[
Y = AX
\]

where \( X \triangleq (x_1, \ldots, x_N)^T \), \( Y \triangleq (y_1, \ldots, y_N)^T \) and \( T \) denotes the transpose of a vector. The transform coefficients are quantized to give \( \hat{Y} \) which is binary coded and transmitted to the receiver. \( \hat{Y} \) is the original vector \( Y \) corrupted with the quantization noise \( Q_N \) introduced during the quantization process. Under ideal channel conditions, the recovered vector is \( \hat{Y} \) with no additional distortion. \( \hat{Y} \) is transformed back into a time-domain vector \( \tilde{X} \) using the transformation matrix \( A^{-1} \) where \( (\cdot)^{-1} \) denotes the inverse of \( \cdot \).

Historically, Transform Coding was used extensively for the coding of picture signals (84,85). Zelinski and Noll also demonstrated its advantage over DPCM for the coding of speech signals (71,72). The success of TC resides in the following two main reasons:

1. The variances of transform coefficients are different with some coefficients having larger variances and others having smaller variances. Data compression is thus achieved by packing the maximum amount of information into the minimum number of coefficients through the transformation process.

2. Since the variances of the transform coefficients are not uniformly distributed, the quantization of the coefficients with larger variances can be performed with more bits than those
with smaller variances and thus a more efficient employment of bits for the quantization process is achieved.

Clearly, the kind of transformation and the bit allocation algorithm determine the performance of a transform coder. There are numerous orthogonal transformations, or more generally, unitary transformations, which can be employed for the coding of speech \((71, 72, 86)\). Among them, Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) are of special interest for speech as the transform coefficients produced correspond to the frequency-domain components of the input signal. Therefore, they can be viewed as a different means to achieve band-division on a block basis. As the blocksize of the transformation used in Transform Coding of speech is usually larger than 128, they are implicitly a narrowband analysis technique. Also, working in the frequency domain enables the coder to achieve noise spectral shaping through an appropriate bit allocation algorithm \((67)\).

Figure 2.3.14 illustrates the system block-diagram of an Adaptive Transform Coding (ATC) scheme. A block of \(N\) sample speech is first buffered and normalized before being transformed. The transform coefficients are then quantized with different number of bits and with different quantizer step-sizes. The bit allocation pattern and the quantizer step-sizes are obtained from a coarse description of the short-time transform output. At the receiver, inverse transformation of the received samples yields the recovered speech.
Figure 2.3.13 Block diagram of sub-band coding (SBC) (after Reference 87)

Figure 2.3.14 Block diagram of adaptive transform coding (ATC) (after Reference 87)
The encoding of narrowband speech using ATC scheme at 16 kbps and above was reported to provide excellent quality recovered speech \(^{(71)}\). Later, however, Galand reported \(^{(82)}\) the problem of inter-block discontinuities at 2 bits per Nyquist sample. Below the bit rate of 16 kbps, the quality of the recovered speech deteriorates rapidly as a lowpass effect and a 'burbly' distortion become evident. Chapter 6 describes the technique of ATC in greater detail.

2.3.3.3 Phase Vocoder (\(\Phi V\))

The Phase Vocoder \(^{(88,14)}\), developed by Flanagan, divides the full band speech spectrum into 30 subbands using a bank of contiguous bandpass filters. The amplitudes of the subband signals are digitally encoded and transmitted to the receiver. Bandwidth saving is accomplished through lowpass filtering the slowly varying phase derivative and amplitude signals and thus less bits are required for their encoding.

The system block diagram of a phase vocoder is shown in Figure 2.3.15. The input speech signal \(x(t)\) is passed through a parallel bank of contiguous bandpass filters. Let the impulse response of the \(n\)th filter be

\[
g_n(t) = h(t) \cos \omega_n t
\]

where \(h(t)\) is the impulse response of a physically-realizable lowpass filter and \(\omega_n\) is the centre frequency of the \(n\)th frequency band. The output signal \(x_n(t)\) of the \(n\)th filter is therefore given by:
Figure 2.3.15a Speech synthesis from short-time amplitude and phase-derivative spectra (after reference 14)

Figure 2.3.15b Programmed analysis operations for the phase vocoder (after reference 14)
\[ x_n(t) = \int_{-\infty}^{t} x(\lambda) h(t - \lambda) \cos[j \omega_n (t - \lambda)] d\lambda \]

\[ = \text{Re} \left \{ \exp(j \omega_n t) \int_{-\infty}^{t} x(\lambda) h(t - \lambda) \exp(-j \omega_n \lambda) d\lambda \right \} \]

(2.3.49)

Therefore, \( x_n(t) \) is the Fourier transform of that part of \( x(t) \) 'viewed' through the sliding time aperture \( h(t) \) and evaluated at the frequency \( \omega_n \). If the complex value of the integral in equation (2.3.49) is denoted as \( X(\omega_n, t) \), its magnitude is the short-time amplitude spectrum \( |X(\omega_n, t)| \) and its angle is the short-time phase spectrum \( \phi(\omega_n, t) \).

\( x_n(t) \) is then given by

\[ x_n(t) = \text{Re} \left \{ \exp(j \omega_n t) X(\omega_n, t) \right \} \]

or

\[ x_n(t) = |X(\omega_n, t)| \cos \left \{ \omega_n t + \phi(\omega_n, t) \right \} \]  

(2.3.50)

Each \( x_n(t) \) may therefore be described as the simultaneous amplitude and phase modulation of a carrier \( \cos \omega_n t \) by the short-time amplitude and phase spectra of \( x(t) \), both evaluated at \( \omega_n \).

To achieve bandwidth saving, the relatively well behaved amplitude spectrum \( |X(\omega_n, t)| \) is lowpass filtered to about 30 Hz before being digitally encoded. A similar procedure is applied to the phase derivatives \( \dot{\phi}(\omega_n, t) \). At the receiver, the phase spectrum is recovered to within an additive constant, by integrating the values of the derivatives.
The reconstruction of the original speech signal is accomplished by summing the outputs of the \( n \) oscillators modulated in phase and amplitude by the received spectra, i.e.

\[
\tilde{x}(t) = \sum_{i=1}^{n} \tilde{x}_i(t),
\]

(2.3.51)

where

\[
\tilde{x}_i(t) = |X(\omega_i, t)|_q \cos (\omega_i + \tilde{\phi}(\omega_i, t)),
\]

(2.3.52)

\[
\tilde{\phi}(\omega_i, t) = \int_{0}^{t} \tilde{\phi}_q(\omega_i, t) dt
\]

(2.3.53)

and \( |X(\omega_i, t)|_q, \tilde{\phi}_q(\omega_i, t) \) represent the quantized values of the amplitude and phase spectra respectively.

It was reported\(^{(15)}\) that computer simulations of phase vocoder produced good quality narrowband speech at the bit rates of 9.6 and 7.2 kbps. The disadvantage of this scheme is of course the great processing requirements involved.

2.3.4 Hybrid Coding

The waveform coding and vocoding techniques described in the previous sections represent two opposing classes of speech encoding methods. The former offers a relatively higher quality recovered speech at a higher transmission bit rate and with a lower system complexity as opposed to the lower quality, lower bit rate and higher system complexity
of the latter. It is therefore a very natural development for the third category of coding method, Hybrid Coding (HC), to combine the features of both the other two techniques in order to achieve a performance, bit rate and system complexity in between the two extremes.

In an attempt to improve the quality of the LPC vocoder speech, Un and Magill (89) proposed the concept of residual-excited linear prediction (RELP) vocoder. In the RELP system, vocal tract modelling is done by the LPC technique, and a modification of the LPC residual signal is used as the excitation signal. After lowpass filtering, the residual signal is coded by adaptive delta modulation and is spectrally flattened before being fed, at the receiver, into the LPC synthesizer. The range of the transmission rate is typically between 6 and 9.6 kbps. As no pitch extraction is required, RELP is easier to implement than the conventional LPC vocoder. To further reduce the complexity of RELP, Viswanathan (90) suggested the derivation of the excitation signal from the original signal rather than from the residual signal. Thus the filtering process at the transmitter is avoided. This modified version of RELP is termed the voice-excited linear prediction (VELP) vocoder.

In a second technique called harmonic scaling, the speech signal is parametrically compressed in bandwidth and sampling rate. The compressed signal is then digitally encoded by waveform coding methods. Bit rate reduction is thus achieved. Harmonic scaling can be implemented.
either in the time or frequency-domain. The time-domain technique requires explicit pitch tracking. Once the pitch is known, the time-domain operations are simple and result in good quality scaled and reconstructed speech\(^{(91)}\). Frequency-domain scaling techniques are generally much more complex but do not require explicit pitch tracking\(^{(92-94)}\).

The various techniques of harmonic scaling have been shown to provide high quality narrowband speech at the bit rates of 16 kbps and below.

2.3.4.1 Residual-Excited Linear Prediction (RELP) Vocoder\(^{(89)}\)

Conventional LPC systems model the vocal tract, over a short-time segment of about 20 msec, by a time-invariant all-pole recursive digital filter. The excitation source is represented by a periodic pulse train with a period equal to the pitch period of the original speech or by a white noise source. In the residual LPC vocoder (RELP) (Figure 2.3.16), the vocal tract is characterized in the same way as in the pitch-excited LPC vocoder. However, the excitation source is derived from the residual signal which is lowpass filtered to 800 Hz and digitally encoded for transmission. At the receiver, the lowpass filtered residual signal is spectrally flattened to recover the high-frequency harmonics. The flattened signal is usually mixed with an appropriate amount of white noise so that the synthesized speech is perceptually more pleasing. In principle, RELP is very similar to ADPCM and APC described in the waveform coding section. The two waveform coders differ from RELP in that the lowpass filtering,
Figure 2.3.16  System block diagram of a Residual-Excited Linear Prediction (RELP) Vocoder (a) Transmitter (b) Receiver (after Reference 89)
decimation and spectral flattening processes are not employed to achieve bandwidth reduction. The other difference is that the waveform coders are able to reconstruct the speech waveforms to a very accurate degree, when given sufficient number of bits for the quantization of the error sequence, which is not the case in RELP no matter how fine the quantization process is. Signal distortions are unavoidable even under no quantization condition due to the various compression processes involved.

(i) Spectral Flattening

Spectral Flattening is a salient feature in the RELP and VELP schemes. The quality and naturalness of these coders depend to an extent on the type of spectral flattener used. Spectral flattening is usually achieved through non-linear distortion. One of the non-linear distortion methods used to increase the mean rate of zero-crossings consists of a piece-wise linear network with an input-output characteristic of straight line segments shown in Figure 2.3.17a, which resembles the character 'W'. The 'W' function spectral flattener\(^{(95)}\) spreads the spectrum more for large input amplitudes than for small input amplitudes. In order to overcome this effect of different input amplitudes, instantaneous logarithmic compression is applied preceding the 'W' function. The combined characteristic is shown in Figure 2.3.17b.

Rectification is another commonly used non-linear distortion method to achieve spectral flattening\(^{(96-101)}\). In general, a rectifier
Figure 2.3.17(a) The input-output characteristics of a non-linear 'W' network

Figure 2.3.17(b) The combined input-output characteristics of (a) and a logarithmic compressor (after reference 95)
operating on a signal $x(t)$ has the following input/output characteristics (97):

$$y(t) = \frac{1}{2} \{ (1 + \alpha) |x(t)| + (1 - \alpha)x(t) \} ;$$

$$0 \leq x \leq 1$$

(2.3.54)

where $| \cdot |$ denotes absolute value and $\alpha$ represents the extent of rectification. $\alpha = 0$ gives half-wave rectification and $\alpha = 1$ corresponds to full-wave rectification. The value of $\alpha$ can be adjusted so that the synthesized speech is subjectively more pleasing.

Besides the method of non-linear distortion, high frequency regeneration could be achieved through spectral duplication. Spectral duplication can take the form of (a) spectral folding and (b) spectral translation (97). To perform an $L$ band spectral fold, one simply inserts $L-1$ zeros between samples of the transmitted baseband signal. Figure 2.3.18b shows the result of 3 band spectral folding. The process of integer-band spectral translation is illustrated in Figure 2.3.19. The multiplication by $(-1)^t$ produces a signal which has a mirror image spectrum of the signal before multiplication. Again, upsampling by inserting zeros is employed to achieve spectral folding. The multiple bandpass filter $H(z)$ passes those bands that have the same shape as the baseband, i.e. every other band, and the filter $1 - H(z)$ passes the intervening bands. The sum of the outputs of the two filters constitutes the required extended spectrum. Figure 2.3.18c shows the result of spectral translation.
Figure 2.3.18  
(a) Baseband spectrum  
(b) 3-band spectral folding  
(c) 3-band spectral translation  
(after Reference 97)

![Diagram](image)

Figure 2.3.19  Receiver for baseband coder that uses integer-band spectral translation  
(after Reference 97)
The use of spectral duplication for high frequency regeneration is found to produce low level background tones which is different from the 'roughness' produced by the rectification method. Makhoul and Berouti (97) suggested an improved scheme which seeks to preserve the harmonic structure of the baseband in order to eliminate the undesired background tones. This is done by adjusting the width of the baseband spectrum to be a multiple of the short-term pitch fundamental frequency.

2.3.4.2 Frequency Scaling of Speech Signals

Frequency scaling is a useful method to reduce the bandwidth of the input speech signal by half or more, and thus reducing the sampling frequency, so that the transmission bit rate can be substantially reduced. A general block diagram of a digital system using frequency scaling as a preprocessing technique (92) is shown in Figure 2.3.20. In this Figure, the input speech signal (voiced speech) is shown to have a spectral envelope with formant peaks and an underlying fine pitch-harmonics structure. The compressed signal is a frequency-scaled version of the input signal. Frequency scaling can be carried out either in the time or frequency-domain. The resultant distortion after the compression and expansion processes is dependent on the scaling algorithm used. The waveform coding process in between them also determines the quality of the recovered speech at the output of the system. The various techniques of frequency scaling can be described in the general
Figure 2.3.20 General block diagram of a digital coding system which applies frequency scaling for bit rate reduction (after Reference 92)
framework of short-time Fourier transform (STFT) analysis, modification and synthesis (92).

The relation between STFT and filter bank analysis is well established (102-104). Consider a filter bank with N contiguous bandpass filters. Let \( h_k(t) \), the impulse response of the \( k \)th filter, be given by:

\[
h_k(t) = v_k(t)e^{j\omega_k t}
\]

where \( v_k(t) \) is the impulse response of the lowpass prototype of \( h_k(t) \) and \( \omega_k \) is the centre frequency of that filter. The output signal of the \( k \)th filter has the general form

\[
z_k(t) = A_k(t)e^{j\phi_k(t)}
\]

\[
= A_k(t)e^{j(\omega_k t + \phi_k(t))}
\]

where \( A_k(t) \) is the amplitude function and \( \phi_k(t) \) is the total phase function which consists of the components \( \omega_k t \) and \( \phi_k(t) \). The signal \( z_k(t) \) can therefore be interpreted as being the result of simultaneous modulation of the amplitude and phase of the carrier \( e^{j\omega_k t} \) by the amplitude and phase signal \( A_k(t) \) and \( \phi_k(t) \).

Since \( z_k(t) \) is the output signal from a bandpass filter having an impulse response \( h_k(t) \), it can be expressed as the convolution between the input speech signal \( x(t) \) and \( h_k(t) \), i.e.
\[ z_k(t) = \int_{-\infty}^{\infty} x(\tau) \omega_k(t - \tau) e^{j\omega_k(t - \tau)} d\tau \]
\[ = e^{j\omega_k t} X(\omega_k, t) \quad (2.3.57) \]

where
\[ X(\omega_k, t) = \int_{-\infty}^{\infty} x(\tau) \omega_k(t - \tau) e^{-j\omega_k \tau} d\tau \quad (2.3.58) \]

Comparing Equations 2.3.56 and 2.3.57, we have
\[ X(\omega_k, t) = A_k(t) e^{j\phi_k(t)} \quad (2.3.59) \]

The expression for \( X(\omega_k, t) \) in Equation 2.3.58 shows that \( X(\omega_k, t) \) is equal to the value of the short-time Fourier transform of \( x(t) \) weighted by the window function \( \omega_k(t) \) and evaluated at the frequency \( \omega = \omega_k \).

The next step to achieve frequency scaling of the output signal \( z_k(t) \) from the bandpass filter involves the modification of its STFT \( X(\omega_k, t) \).

Let \( z_{qk}(t) \) denote the frequency scaled version of \( z_k(t) \) and the modified STFT of \( X(\omega_k, t) \) be \( X_q(\omega_k, t) \). We then have
\[ z_{qk}(t) = X_q(\omega_k, t) e^{j\omega_k t} \quad (2.3.60) \]

In general, the modification of the STFT, \( X(\omega_k, t) \) only achieves partial scaling and therefore additional filter, \( \omega_{qk}(t) \), having a bandwidth \( q \) times that of \( \omega_k(t) \), is required to prevent excessive inter-band aliasing when the partially scaled subband signals are combined. The
filters \( w_k(t) \) and \( w_{qk}(t) \) are generally known as the analysis and synthesis filters. Figure 2.3.21 shows a general block diagram for frequency scaling which is based on modifying the STFT of the input signal. The output scaled signal, \( y_q(t) \), of the system is the sum of the individual scaled and filtered subband signals \( \hat{z}_{qk}(t) \), i.e.

\[
y_q(t) = \sum_{i=1}^{N} \hat{z}_{qk}(t) \tag{2.3.61}
\]

There are several known frequency scaling techniques which can be described in terms of the corresponding STFT modifications they achieved. The main features of each technique are described in the following sections.

(i) Analytic Signal Rooting Technique

This technique has the following main features:

(1) The number of bandpass filters is chosen to match the formant structure of the speech signal so that no more than one formant is present in each subband.

(2) The scaling of \( z_k(t) \) to \( z_{qk}(t) \) is achieved simply by raising the power of \( z_k(t) \) to \( q \). In the context of STFT modification, the scaling operation can be shown to be equivalent to:

\[
X_q(\omega_k,t) = \left[ X(\omega_k,t) \right]^q \tag{2.3.62}
\]

In terms of the modulating amplitude and phase signal this modification corresponds to
Figure 2.3.21 General block diagram of a frequency scaling system based on short-time spectral modification (after Reference 92)
\[ A_{qk}(t) = \left( A_k(t) \right)^q \quad (2.3.63a) \]

\[ \phi_{qk}(t) = q \phi_k(t) \quad (2.3.63b) \]

The analysis done in references (106, 107) shows that

(1) The phase scaling operation in (2.3.63b) scales the instantaneous frequency of the dominant harmonic in each band by \( q \), but the lower amplitude harmonics are shifted in such a way that their spacing from the dominant harmonic remains unchanged.

(2) The amplitude scaling operation in (2.3.63a) scales the magnitude of the non-dominant harmonics relative to the amplitude of the dominant component.

Obviously, this technique does not achieve perfect frequency scaling. However, the algorithm is relatively easy to implement.

(ii) Phase Vocoder Technique \(^{(88)}\)

This technique has the following main features:

(1) A relatively large number of bandpass filters are used to match the harmonic structure of voiced speech so that, preferably, no more than one pitch harmonic is present in each subband.

(2) The individual harmonics are separately scaled. Let \( \Omega_k \) be the pitch harmonic frequency in the kth subband (with centre frequency
where \( \dot{\hat{\psi}}(t) \) denotes the contribution of the phase variations to the bandwidth of the pitch harmonic in the kth subband. The phase vocoder technique scales the phase derivative by \( q \) but it leaves the amplitude signal unmodified. The scaling operation can be expressed by:

\[
A_{qk}(t) = A_k(t) \quad (2.3.65a)
\]

\[
\phi_{qk}(t) = \int_{t_o}^{t} q \dot{\phi}_k(t) dt \quad (2.3.65b)
\]

It should be noted that the constant phase \( \phi_k(t_0) \) is discarded and it can have an effect on the shape of the scaled signal waveform.

The analytic signal rooting technique and the phase vocoder technique described so far are frequency-domain scaling techniques. They are generally known by their individual names. The Frequency Domain Harmonic Scaling (FDHS) techniques to be described in the following section refers specifically to the scheme suggested by Malah and Flanagan (92).

(iii) Frequency Domain Harmonic Scaling (FDHS) Technique

As in the phase vocoder, the FDHS technique aims at scaling the individual pitch harmonics of voiced speech signal. It differs
from the phase vocoder in that the total phase, including the constant phase term, is scaled. As in the phase vocoder, the amplitude signals remain unmodified. The modified STFT amplitude and phase components are given by

\[ A_{qk}(t) = A_k(t) \quad (2.3.66a) \]
\[ \phi_{qk}(t) = q \phi_k(t) \quad (2.3.66b) \]

For non-integer values of \( q \), the phase modification in (2.3.66b) must be performed on the unwrapped phase in order to avoid incorrect phase modification due to the phase ambiguity of multiple of \( 2\pi \). However, for a particular value of \( q = \frac{1}{2} \), phase modification can be achieved through a simple sign-tracking algorithm which avoids the need for explicit phase computation and unwrapping (93).

(iv) Time Domain Harmonic Scaling (TDHS)

The approach taken by the TDHS technique is to incorporate pitch information (obtained by a separate pitch detector) into the scaling process. Once the pitch period is known, a very elegant time-domain algorithm can be used to achieve frequency compression and expansion of the signal (91). For a compression and expansion by a factor of two, the algorithm can be described as follows (see Figure 2.3.22):
The input speech \( x(n) \) is pitch synchronously divided into blocks of \( 2p \) samples, where \( p \) is the measured pitch period. The first block of \( p \) samples is weighted by a window \( w(n) \) which linearly decreases from 1 to 0 across the block. The second block of \( p \) samples is weighted by a window \( 1 - w(n) \). These two blocks of weighted samples are then added to produce the \( p \) frequency scaled samples \( x_c(n) \), which looks mostly like the first block of \( x(n) \) at its beginning and mostly like the second block of \( x(n) \) at the end. In this way, the concatenation of the blocks of \( x_c(n) \) forms a continuous waveform without the end effects from block to block.

The expression process of the TDHS algorithm is depicted in Figure 2.3.22b. In this case, \( 3p \) compressed samples \( x_c(n) \) are used to form the \( 2p \) expanded samples \( \hat{x}(n) \) by the following procedure: The first block of \( 2p \) compressed samples \( x_c(n) \) are weighted by the window \( 1 - w(n) \) which linearly increases from 0 to 1 across the block. The second block of \( 2p \) \( x_c(n) \) samples, which includes the last \( p \) samples of the first block, is weighted similarly by the window \( w(n) \). These two blocks of weighted samples are then added to produce the \( 2p \) recovered samples \( \hat{x}(n) \). Next, the windows are shifted by \( p \) samples and the processes of windowing and summation are repeated. Thus, for every \( p \) samples of the compressed samples \( x_c(n) \), \( 2p \) samples of the expanded samples \( \hat{x}(n) \) are produced.
Figure 2.3.22  Illustration of Time Domain Harmonic Scaling
(a) Compression,  (b) Expansion
(after Reference 91)
(v) The TDHS/(SBC, ATC) Speech Coders

The elegant TDHS can be used in conjunction with the frequency-domain waveform coder like SBC or ATC to achieve a bit rate of 8 to 9.6 kbps for an improved quality narrowband speech. The TDHS algorithm has the attractive feature that the implementations of the pitch extractor and the scaling algorithm can be modularized and separately realized by DSP chips. Good quality recovered speech both in computer simulation and real-time implementation had been reported in the literature \(^{(108, 112)}\). Figure 2.3.23 shows the system block diagram of the TDHS/SBC speech coder.

2.4 ASSESSMENT OF CODER PERFORMANCE

The assessment of the performance of speech coders relies on objective measurement like signal-to-quantization noise (SNR) ratio and also on subjective listening tests. The use of SNR as a performance indicator is normally applicable to waveform coders which employ the various time and frequency-domain coding techniques to preserve the speech waveform as accurately as possible. Though it is widely used in assessing coder performance, it is by no means an absolute and ideal indicator. More often than not, it serves as a guidepost for coder design and provides a means for objective comparisons between similar coding techniques. Its importance
Figure 2.3.23 Block diagram of the TDHS/SBC coder for 9.6 kbits/s (after Reference 11)
diminishes as the bit rate of a coder drops down to 2 bits/Nyquist
sample and below. At the bit rate of around 1.5 to 2 bits/Nyquist
sample, coder design criteria invariably differ from the minimum
quantization noise power criterion used for the design of high bit
rate coder. At this range of transmission bit rate, the spectral
shaping of the output noise of the time or frequency-domain coders
to achieve optimum perceptual performance always leads to lower SNR
measurements compared to the similar coders which are designed to
maximize SNR performance. In the category of analysis-synthesis
coding or vocoding techniques, SNR measurement becomes almost irrelevant
and coders' performance is assessed by subjective tests which focus on
intelligibility and recognizability. The methodology of the intelligibility and recognizability testing procedures, as well as the statistical
techniques used to interpret the results, are subjects beyond the scope
of this thesis.

The two objective performance indicators employed throughout the
research are the average segmental signal-to-quantization noise ratio
(SNRSEG)(113,114) and the long-term average power spectral density
plots of the output noise of a coder. The latter is proposed as a
means to indicate the long-term average frequency distribution of the
output noise power of a coder. When plotted together with the long-
term average power spectral density of speech, it serves as a very
good indication of the spectral relation between the output noise
and speech. It can also be used to check the computer simulation results and helps to interpret the perceptual quality of the recovered speech. In the course of the research, it was found to be a good objective indicator to supplement the more conventional SNRSEG measurement.

2.4.1 Segmental Signal-to-Quantization Noise Ratio (SNRSEG)\(^{(113,114)}\)

Segmental signal-to-quantization noise ratio (SNRSEG) is obtained by averaging block SNR values measured every 16 msec intervals for usually a sentence of speech input. SNRSEG can be expressed as:

\[
\text{SNRSEG}(\text{dB}) = \frac{1}{NN} \sum_{m=1}^{NN} 10 \log_{10} \frac{W \sum_{i=1}^{W} x_i^2}{W \sum_{i=1}^{W} (x_i + (m-1)w - \hat{x}_i + (m-1)\hat{w})^2}
\]

(2.4.1)

where \(NN\) is the total number of blocks in the speech sentence under measurement, \(x_i\) and \(\hat{x}_i\) are the \(i\)th sample of the input speech and the recovered speech respectively and \(w\) is the number of samples in a block of typically 16 to 20 msec. One important feature in this measurement is the exclusion of silence blocks in the speech input data under measurement. Block SNR measurements for silence periods usually have very high values and their inclusion would distort the true performance assessment. Thus an energy threshold detector is used to detect the presence of the silence blocks to avoid their inclusion in the averaging process.
Throughout the course of the research SNRSEG was found to be speaker and sentence dependent. Therefore, it is more suitable for comparison purposes rather than to be taken as an absolute performance indicator. Another form of SNR measurement is the long-term average SNR ratio (4) defined as

\[
\text{SNR} = 10 \log_{10} \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \hat{x}_i)^2}
\]

where \(n\) is the total number of samples in the whole speech sentence under measurement. Therefore, it is also termed total SNR. As it was found to be less accurate as a performance indicator than SNRSEG (20), it was not used in this research project. Hereafter, the term SNR will simply mean SNRSEG described above.

2.4.2 Long-term Average Power Spectral Density of Output Noise

The second objective measurement of the performance of a coder is the long-term average power spectral density of the output noise (115). Output noise samples are obtained by subtracting the recovered speech samples from the original speech samples, taking into account the total coder delay. To obtain the long-term average noise spectrum, \(N_{S, L.T.}\), the short-term noise spectrum, \(N_{S, S.T.}\), is calculated for every block of 16 msec (256 samples) with each block overlapping half of the previous block as shown below:
The samples are Hanning windowed, before the FFT transform, to reduce the effect of rippling\(^{(116)}\). The magnitudes of the frequency components are then squared and converted into dB to form the short-term power spectrum. The final long-term average power spectrum of the output noise is obtained by averaging the short-term spectra, i.e.

\[
NS_{L.T.} = \frac{1}{2\cdot NN-1} \sum_{i=1}^{2\cdot NN-1} NS_{(S.T.)i}
\]  

(2.4.3)

where \(NN\) is the total number of blocks in a sentence.

2.4.3 Informal Subjective Listening Tests

The assessment of the performance of a speech coder is incomplete without some form of subjective listening tests. Though formal listening tests are probably the most reliable indicators of the performance of speech coders, they are not usually resorted to because of the complexity, difficulties and facilities involved in carrying out the tests.

To provide a subjective assessment of the performance of the coders investigated in this research programme, informal listening tests were employed instead. Informal listening tests
usually involve a number of both experienced and inexperienced listeners. They are presented with pairs of speech sentences, either the original and the recovered speech or two recovered speeches from different coders, for their comparisons and comments. Their preferences and qualitative descriptions of the distortions, if any, in the recovered speech are then recorded as a form of subjective assessment of the coders' performance.

2.5 HARDWARE ISSUES AND CODER COMPLEXITY

2.5.1 Hardware Issues

Recent advances in VLSI technology and the advent of digital signal processing devices have enabled the real-time implementation of complex speech coding algorithms. Thus speech coding research no more relies solely on computer simulation though it is still the most efficient tool for investigation. Indeed, a considerable amount of research and development on the real-time implementations of promising speech coding algorithms have been done by the various telecommunications network operators and manufacturers. Generally, digital signal processing devices used in speech coding can be divided into two types:

(1) Custom Chips and Devices

These are chips and devices that are specifically designed and constructed to perform a specific coding algorithm in real-time. Examples in the waveform coding category are chips for μ-law
A/D and D/A conversion and adaptive delta-modulation. In the area of vocoding, one of the earliest devices available on the market is the Texas Instrument's 'speak and spell' chip. The prohibitive cost of developing these devices has restricted their use up to now, to the more commercially acceptable and generally less complex coding techniques.

(2) High Speed Microprocessors and Programmable ICs

The second group of devices includes high speed microprocessors and programmable ICs that have been modified to become more suitable for signal processing, particularly in real-time. Notable examples are the Bell Laboratories' Digital Signal Processing (DSP) IC (111,117), the NEC 7720 DSP chip (118) and the Texas Instruments' TMS 320 DSP chip (119). They can be used to implement in real-time coding techniques as complex as LPC and formant vocoder.

Both of these types of devices will undoubtedly have great influence in the design of the future telecommunication networks.

2.5.2 Coder Complexity

Complexity is obviously an important issue in coder design, since it is invariably tied up with implementability and cost. The complexity of a coding scheme is perhaps the most complicated variable to try to assess. Generally, the complexity of a coder is assessed in terms
of the amount of processing involved, like the number of additions/subtractions and multiplication/divisions, and the memory size required to store the programme instructions and the intermediate value of the variables. Alternatively, it can be gauged in terms of the number of chips and DSP devices required for real-time implementation. Flanagan, et al. (15) provided an approximate ranking of a number of speech coders in terms of complexity, by comparing each to the simple adaptive delta modulator (ADM), which is assigned a complexity factor of unity. (See Table 2.5.1).
<table>
<thead>
<tr>
<th>Relative complexity*</th>
<th>Coder</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>ADM (adaptive delta modulator)</td>
</tr>
<tr>
<td>1</td>
<td>ADPCM (adaptive differential PCM)</td>
</tr>
<tr>
<td>5</td>
<td>SBC (subband coder)</td>
</tr>
<tr>
<td>5</td>
<td>P-P ADPCM (ADPCM with pitch prediction)</td>
</tr>
<tr>
<td>50</td>
<td>APC (adaptive predictive coder)</td>
</tr>
<tr>
<td>50</td>
<td>ATC (adaptive transform coder)</td>
</tr>
<tr>
<td>50</td>
<td>V (phase vocoder)</td>
</tr>
<tr>
<td>50</td>
<td>VEV (voice-excited vocoder)</td>
</tr>
<tr>
<td>100</td>
<td>LPC (linear-predictive coding vocoder)</td>
</tr>
<tr>
<td>100</td>
<td>CV (channel vocoder)</td>
</tr>
<tr>
<td>200</td>
<td>ORTHOG (LPC vocoder with orthogonalized coefficients)</td>
</tr>
<tr>
<td>500</td>
<td>FORMANT (formant vocoder)</td>
</tr>
<tr>
<td>1000</td>
<td>ARTICULATORY (vocal tract synthesizer; synthesis from printed text)</td>
</tr>
</tbody>
</table>

* Essentially a relative count of logic gates. These numbers are very approximate, and depend upon circuit architecture.

Table 2.5.1 Relative Complexity of the Various Speech Coding Algorithms
(after Reference 15)
2.6 CONCLUSION

The various techniques and issues related to the design of a speech coding system, presented in this chapter, are at best a very shallow introduction to the enormous wealth of information and knowledge on the subject accumulated in the past few decades or so. The design of speech coding systems is indeed a complicated and challenging task, that involves careful selection of a coding algorithm to suit the individual requirements, like quality and bit rate, under specific constraints like complexity, cost and coder delay. With the availability of low cost, high speed and simple to use digital signal processing devices, more complicated and efficient speech coding techniques are increasingly being employed and integrated into the future generation of telecommunication networks for the benefit of mankind. Together with the equally important and large amount of work done in the parallel areas of speech synthesis and recognition, though not discussed in this thesis, one is left without any doubt that a speech technology revolution is taking place!
CHAPTER 3

TIME AND FREQUENCY DOMAIN CODING OF

WIDEBAND SPEECH AT 64 and 56 KBPS
3.1 INTRODUCTION

The existing telephone networks restrict speech signals to a narrow bandwidth of 0.3 - 3.4 kHz and employ either A-law or μ-law Pulse Code Modulation (PCM) for their digital transmission at 64 kbps. This arrangement is increasingly under review on the aspects of quality and efficiency by the Consultative Committee for International Telephone and Telegraph (CCITT). As the frequency range of unvoiced speech sounds extend beyond the 3.4 kHz limit, the imposed bandwidth limitation causes attenuation and distortion thus resulting in loss of speech intelligibility and quality. To improve the quality of voice services over the emerging Integrated Services Digital Networks (ISDN), the bandwidth of speech signals obviously has to be increased. The bandwidth of 0-7 kHz has been accepted by CCITT as the minimum for wideband or commentary grade speech signals. In fact, coding of wideband speech at 64 kbps or below is currently being studied by the CCITT Rapporteur's Group SGXVIII (120-122). The other area that receives high priority examination by CCITT is the 32 kbps coding of narrowband speech using more efficient algorithms than PCM (124).

In this chapter, a computer simulation study of the various simple time and frequency domain wideband speech coding techniques at 64 kbps is presented. As the sampling frequency is 16 kHz, the coding system is operating at the relatively high number of 4 bits/sample. Therefore
the coding algorithms which can produce the desired quality of received speech are not expected to have the order of complexity as that of hybrid coding or vocoding systems.

It is well known that differential PCM (DPCM) coders can reduce substantially the transmission bit rate, whilst maintaining toll quality speech, by exploiting the redundancies present in speech. DPCM systems employing an adaptive quantizer or predictor or both are generally known as ADPCM. These coders show considerable improvements in signal to quantization noise ratio and subjective perceptual quality over DPCM. A particular ADPCM system employing the one-word memory Jayant quantizer (AQJ) in the feedforward path and a fixed predictor in the feedback path offers better performance than PCM at the same transmission bit rate with minimal system complexity. It is therefore chosen as the first scheme for investigation. Computer simulation of the system was carried out and its performance evaluated in terms of objective measurements like average segmental SNR value and power spectral density plot of the quantization noise. Informal listening tests were also carried out to give a subjective assessment of the quality of the decoded speech.

To further improve the performance of an ADPCM system, there are in general two different approaches one can adopt. The first approach is to minimize the power of the quantization noise, i.e. to improve the SNR performance, by improving the performance of the quantizer or
predictor or both. The Pitch Compensating Quantizer (125-126) and the Dynamic Ratio Quantizer (127,128) are two possible schemes that offer better adaptive quantization strategies than AQJ. The choice of adaptive prediction is more diversified (37-54, 125-130). There are forward block adaptive predictors (30, 129), Sequential Gradient Estimation Predictor (37,40,41,129), Kalman Adaptive Predictor (129), adaptive Lattice Predictor (51-54), Pitch Predictor (130) etc. All these predictors seek to minimize the variance of the error sequence to be quantized.

The other way to improve the quality of the recovered speech is to reduce the perceptual effect of the quantization noise. Noise shaping (131-139) and subband coding (68-70,73-83) are two possible techniques currently being pursued by many speech coding researchers. Noise shaping involves the use of a feedback filter or pre-filter to alter the spectral shape of otherwise relatively flat quantization noise spectrum. The quantization noise is shaped such that the auditory masking effect of the ear can be exploited.

Subband coding involves the partitioning of speech into frequency subbands followed by separate encoding of the signals in each band. By exploiting the different properties of speech in different bands and allocating bits among subbands according to perceptual palatability, the quantization noise can be shaped and its effect reduced. Further detailed discussion on noise shaping and subband coding techniques are given in the subsequent sections.
In our experiments, a 6th-order adaptive lattice predictor using the sign product method for updating the PACOR coefficients \(^{140,141}\) was chosen to improve the SNR performance of the ADPCM system employing Jayant's quantizer. Also, in shaping the quantization noise spectrum, the simple noise shaping technique achieved by pre-emphasising the speech input prior to ADPCM encoding and de-emphasising the recovered signals after decoding was investigated. A 2-band subband coding scheme was also examined. The system divides the speech band into two equal subbands using quadrature mirror filters \(^{73}\) and encodes each band separately. The performance of the three coders mentioned above will be compared with that of the simplest ADPCM system employing fixed prediction.
3.2 ADPCM EMPLOYING FIXED PREDICTION

The argument for employing DPCM rather than PCM for the coding of speech signals have been given in Chapter 2. Basically, DPCM exploits the high correlation between the adjacent speech samples by quantizing the difference between the input sample and its estimate. As the output noise of the coder is equal to the quantization noise introduced by the quantizer, more accurate quantization implies higher SNR performance. There are various forms of forward and backward adaptive quantization as discussed in Chapter 2. Backward adaptive quantization seems to be more suitable for the 64 kbps system design as it does not incur additional side-information. The other method to reduce quantization noise, and hence lower output noise, is to reduce the variance of the difference sequence to be quantized, by providing more accurate prediction of the input sequence using adaptive prediction. Again, backward prediction is preferred in general as it does not incur the transmission of side-information.

An ADPCM coder employing fixed prediction and a backward adaptive quantizer such as the one-word memory adaptive Jayant quantizer (AQJ) is perhaps the simplest form of all the ADPCM coders. A schematic block diagram of such a coder is shown in Figure 3.2.1. Analogue speech is lowpass filtered to the bandwidth of 0-7 kHz and sampled at 16 kHz prior to encoding by the ADPCM coder. The adaptive Jayant quantizer in the feedforward path and the spectral envelope predictor in the feedback path are the two main elements in the coder.
Figure 3.2.1 ADPCM with adaptive Jayant quantizer
The predictor is a linear all zero envelope predictor with a transfer function $P(z)$ given by:

$$P(z) = \sum_{i=1}^{N} a_i z^{-i} \quad \text{(3.2.1)}$$

where $a_i$'s are the prediction coefficients and $N$ is the order of the predictor. The output $y_r$ of the linear predictor is an estimate of the input sample $x_r$. $y_r$ can be expressed as the sum of the weighted locally decoded samples $\hat{x}'s$, i.e.

$$y_r = \sum_{i=1}^{N} a_i \hat{x}_{r-i} \quad \text{(3.2.2)}$$

The error sample $e_r$ is therefore equal to

$$e_r = x_r - \left( \sum_{i=1}^{N} a_i \hat{x}_{r-i} \right) \quad \text{(3.2.3)}$$

If the quantization noise is $n_r$, it can be easily shown that

$$\hat{x}_r = x_r + n_r \quad \text{(3.2.4)}$$

and hence

$$e_r = x_r - \left( \sum_{i=1}^{N} a_i x_{r-i} + \sum_{i=1}^{N} a_i n_{r-i} \right) \quad \text{(3.2.5)}$$

Under reasonably accurate quantization conditions, the variance of $n_r$ and the expected value of $n_r x_r$ are negligible compared to the variance of $x_r$. Therefore, from Equation (3.2.5), the variance of $e_r$ can be expressed as
\[ \sigma_e^2 = \mathbb{E}(x_r - \sum_{i=1}^{N} a_i x_{r-i})^2 \]

\[ = \mathbb{E}(x_r^2) - 2 \sum_{i=1}^{N} a_i \mathbb{E}(x_r x_{r-1}) + \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}(x_{r-i} x_{r-j}) \]

(3.2.6)

where \( \mathbb{E}(\cdot) \) means the expected value of \( \cdot \).

Equation (3.2.6) can be expressed more compactly in matrix and vector form as

\[ \sigma_e^2 = \sigma_x^2 - 2 A^T G + A^T G A \]  

(3.2.7)

where the coefficient vector \( A \triangleq \{a_1, a_2, \ldots, a_N\}^T \),

\[ G = \{\psi_1, \psi_2, \ldots, \psi_N\} \] and \( R \) is the matrix given by

\[ R = \begin{bmatrix} \psi_0 & \psi_1 & \psi_2 & \cdots & \psi_{N-1} \\ \psi_{-1} & \psi_0 & \psi_1 & \cdots & \psi_{N-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \psi_{-(N-1)} & \cdots & \psi_1 & \psi_0 \end{bmatrix} \]

(3.2.8)

and \( \psi_i = \mathbb{E}(x_r x_{r+i}) \)  

(3.2.9)
To minimize the value of $\sigma_e^2$, Equation (3.2.7) is partially differentiated with respect to each of the coefficients $a_i$'s and the derivatives are set to zero, i.e.

$$\frac{\partial \sigma_e^2}{\partial A} \bigg|_{A=A_{\text{opt}}} = 0$$  \hspace{1cm} (3.2.10)

The set of optimum prediction coefficients $A_{\text{opt}}$ obtained is then given by

$$A_{\text{opt}} = R^{-1}G$$  \hspace{1cm} (3.2.11)

In general, there are two slightly different solutions to Equation (3.2.11) depending on how the correlation function $\psi_i$ is defined. More details on the formulation of $\psi_i$ are given in Section (3.2.2).

### 3.2.1 Adaptive Jayant Quantizer

The one-word memory adaptive Jayant quantizer (AQJ)(31) is a $B$-bit ($B > 1$) quantizer with its step-size $\Delta_r$ updated at every sampling instant $r$. The output of the quantizer at the time instant $r$ can be expressed as

$$q_r = p_r \cdot \frac{\Delta_r}{2}$$  \hspace{1cm} (3.2.12)

where

$$p_r = \pm 1, \pm 3, \ldots, \pm (2^B - 1)$$  \hspace{1cm} (3.2.13)

and

$$\Delta_r > 0.$$
The step-size has an instantaneous time-invariant adaptation strategy given by

\[ \Delta_r = \Delta_{r-1} M(|P_{r-1}|) \]  

(3.2.14)

where \( M \) is a multiplier function of the various code-word magnitudes \( |P_{r-1}| \). The optimum multiplier functions for \( B = 2, 3, 4 \) and 5 bits for ADPCM coding of narrowband speech signals were obtained by Jayant (31) using a computer search technique. They are given in Table 3.2.1.

| \( B \) | \( |P_{r-1}| \) | \( M(|P_{r-1}|) \) |
|-------|---------|----------------|
| 2     | 1       | 0.8            |
|       | 3       | 1.6            |
| 3     | 1       | 0.9            |
|       | 3       | 0.9            |
|       | 5       | 1.25           |
|       | 7       | 1.75           |
| 4     | 1       | 0.9            |
|       | 3       | 0.9            |
|       | 5       | 0.9            |
|       | 7       | 0.9            |
|       | 9       | 1.2            |
|       | 11      | 1.6            |
|       | 13      | 2.0            |
|       | 15      | 2.4            |
| 5     | 1       | 0.9            |
|       | 3       | 0.9            |
|       | 5       | 0.9            |
|       | 7       | 0.9            |
|       | 9       | 0.95           |
|       | 11      | 0.95           |
|       | 13      | 0.95           |
|       | 15      | 0.95           |
| 6     | 17      | 1.2            |
|       | 19      | 1.5            |
|       | 21      | 1.8            |
|       | 23      | 2.1            |
|       | 25      | 2.4            |
|       | 27      | 2.7            |
|       | 29      | 3.0            |
|       | 31      | 3.3            |

Table 3.2.1 The Jayant multiplier functions for ADPCM coding of speech signals.
The design of AQJ is based on the observation that there are two distinct forms of quantization distortions, namely overload distortion and granular noise. Overload distortion occurs when the quantizer step-size is too small with respect to the input signal. If the quantizer step-size is too big, the quantization error is in the form of granular noise. Therefore AQJ increases the step-size during overload by multiplying the present step-size by a multiplier which is greater than one. As can be seen from Table 3.2.1, all multipliers which correspond to the outer output levels have values greater than one. Conversely, the step-size decreases during granularity by using multipliers which are smaller than one.

3.2.1.1 Robust Jayant Quantizer

Although AQJ performs well under ideal channel condition, it is susceptible to transmission errors. The step-size at a particular time-instant \( r + 1 \) depends on the entire past of the quantizer output, i.e.

\[
\Delta_{r+1} = \prod_{m=0}^{r} M(|P_m|) \Delta_0
\]  

(3.2.15)

where \( \Delta_0 \) is the initial step-size.

Let \( \Delta' \) and \( p' \) be the decoded values of \( \Delta \) and \( p \) at the receiver and assume that at time \( \xi < r \), \( p_{\xi} = i \) while a transmission error causes \( p_{\xi}' = j \). If there is no other error,
An error like this can cause a multiplicative offset between the receiver and transmitter and in principle, can perpetuate indefinitely.

To reduce the effect of transmission errors on AQJ, Goodman\(^{(32)}\) proposed the modification of the adaptation strategy to the following

\[
\Delta_{r+1} = M(p_r) \Delta_r^\beta 
\]

\[
= \prod_{m=0}^{r} M(p_m)^\beta(y-m) \Delta_0^\beta \gamma+1
\]  \hspace{1cm} (3.2.17)

where \(\beta\) is a constant less than one.

If at time \(r\), \(j\) is the received value of \(p\) instead of \(i\),

\[
\frac{\Delta_{i+1}'}{\Delta_{i+1}} = M(p_i) \beta(y-j)
\] \hspace{1cm} (3.2.18)

As \(\beta < 1\), the offset due to each error decays exponentially with time, causing the quantizer to be more robust in the presence of transmission errors than that defined by Equation (3.2.14) where implicitly, \(\beta = 1\).

The value of \(\beta\) can be any value close to but less than unity. A smaller value of \(\beta\) implies faster error dissipation rate but at the same time the coder's SNR performance is degraded.
3.2.2 Determination of Fixed Prediction Coefficients

There are two different methods of determining the fixed prediction coefficient for DPCM coding of speech signals depending on how the matrix \( R \) and vector \( G \) of Equation (3.2.11) are defined. One method is known as the autocorrelation method and the other is the covariance method (25).

3.2.2.1 Autocorrelation Method

In this method, \( \psi_i \) of Equation (3.2.9) is the autocorrelation function of the signal given by

\[
\psi_i = \sum_{r=-\infty}^{+\infty} x_r x_{r+i}
\]  

(3.2.19)

\( \psi_i \) is an even function, i.e. \( \psi_i = \psi_{-i} \), while the matrix \( R \) is real and symmetric with equal diagonal elements. The matrix \( R \) is therefore of the Toeplitz form and it is known as the autocorrelation matrix.

When the prediction coefficients are required to be updated every \( W \) (\( W \gg \) order of the predictor \( N \)) samples in the case of Forward Block Adaptive Prediction, \( \psi_i \) in the autocorrelation method is defined as

\[
\psi_i = \sum_{r=1}^{W-i} x_r x_{r+i}
\]  

(3.2.20)

Again, the matrix \( R \) is a Toeplitz matrix. If the prediction coefficient vector \( A \) is obtained from Equation (3.2.11) with \( R \) being a Toeplitz matrix, the inverse filter \( H(z) \) at the receiver having a transfer function
\[ H(z) = \frac{1}{1 - P(z)} \] (3.2.21)

is stable as the roots of \(1 - P(z)\) lies within the unit circle \(^{(25)}\).

Notice that Equation (3.2.20) implies the windowing of the input samples by a rectangular function of width \(W\). The process of multiplying a signal by a window function is equivalent to the convolution of the frequency response of the window function with the signal spectrum. This results in a 'smearing' effect in the signal spectrum and the degree of smearing depends on both the size and type of the window. As speech signals are generally stationary for the duration of a few pitch periods, the use of Hamming or Hanning windows of a few pitch periods duration provides a reduction in spectral distortion \(^{(26)}\).

### 3.2.2.2 Covariance Method

In this method, the matrix \(R\) and vector \(G\) are defined by

\[
R = \begin{bmatrix}
\phi_{11} & \phi_{12} & \cdots & \phi_{1N} \\
\phi_{21} & \phi_{22} & \cdots & \phi_{2N} \\
\vdots & \vdots & \ddots & \vdots \\
\phi_{N1} & \phi_{N2} & \cdots & \phi_{NN}
\end{bmatrix}
\] (3.2.22)

\[
G = \begin{bmatrix}
\phi_{10} \\
\phi_{20} \\
\vdots \\
\phi_{N0}
\end{bmatrix}
\] (3.2.23)
where $\phi_{ij} = \sum_{r=1-i}^{W-i} x_r x_{r-i-j}$ (3.2.24)

is the covariance function of the signal samples $x_{-(N-1)}, \ldots, x_{-1}$, $x_0, x_1, \ldots, x_W$. The matrix $G$ formed in this way is termed the covariance matrix.

Equation (3.2.24) can be written as

$$\phi_{ij} = \sum_{r=1-i}^{W-i} x_r x_{r+i-j}$$

or

$$\phi_{ij} = \sum_{r=i-j}^{W-j} x_r x_{r+j-i}$$

Hence $\phi_{ij} = \phi_{ji}$ and the matrix $R$ is symmetrical. However, as the diagonal elements are not equal, it is not a Toeplitz matrix.

Therefore the spectral envelope prediction coefficients determined by the covariance method do not guarantee the stability of the inverse filter at the receiver\(^{(25)}\). As the number of samples required for the computation of the covariance function is now increased to $N+W$ compared to $W$ samples used in the autocorrelation method, the spectral envelope prediction coefficients are more accurately determined especially when $W$ is small. Finally, the covariance method is computationally more demanding than the autocorrelation method\(^{(25)}\).

3.2.2.3 Long-term Fixed Prediction Coefficients for Wideband Speech Sampled at 16 kHz

The first to fourth-order fixed prediction coefficients were derived
from 6 sentences of speech of about 15 seconds in total duration. The long-term autocorrelation function $\psi_1$ was obtained from the 15 seconds of speech samples, sampled at 16 kHz, and the autocorrelation method was used to calculate the prediction coefficients. Table 3.2.2 contains the prediction coefficient values of the 1st, 2nd, 3rd and 4th-order predictor.

<table>
<thead>
<tr>
<th>N</th>
<th>$a_1$</th>
<th>$a_2$</th>
<th>$a_3$</th>
<th>$a_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.9880</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1.8869</td>
<td>-0.9113</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.9542</td>
<td>-1.0511</td>
<td>0.0741</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.9491</td>
<td>-0.9760</td>
<td>-0.0657</td>
<td>0.0715</td>
</tr>
</tbody>
</table>

Table 3.2.2 Predictor coefficients for unprocessed wideband speech at 16 kHz sampling rate
3.3 ADPCM EMPLOYING ADAPTIVE LATTICE PREDICTION

The conventional approach to improve the SNR performance of a differential coding scheme is to apply adaptive prediction rather than fixed prediction so that the coder responds to the time-varying characteristics of the input signals. In the past 10 years or so, much effort has been concentrated on the design of adaptive predictions using various different structures like lattice or linear transversal filters and different adaptation strategies like the forward block or backward sequential techniques. Xydeas (30) and Honig (46) carried out detailed comparisons of the performance of the various adaptive predictors in ADPCM coding of narrowband speech. Their results show that the difference in system performance using the various commonly used adaptive algorithms is not significant. Furthermore, it is well known that a difference of a few dB in system performance does not necessarily give rise to any significant perceptual difference in some cases. These suggest that an adaptive predictor which is easy to implement might be the best choice.

The choice of the Forward Block Adaptive Predictor (FBAP) does not seem to be suitable in this case as the algorithm requires the transmission of prediction coefficients as side-information and it also introduces additional delay due to its block processing nature. The other aspect that has to be considered when choosing an adaptive predictor is the stability of the resulting inverse filter under both the ideal and noisy channel conditions. The lattice structure
kind of predictor using certain reflection coefficient updating algorithms has the inherent advantage that the stability of the inverse filter is guaranteed. Furthermore, its implementation can be simplified by employing the conventional sign-product method (140). Therefore, the ADPCM coder employing an adaptive Jayant quantizer and a 6-th order lattice predictor is examined next.

The schematic block diagram of the ADPCM coder employing adaptive lattice prediction is shown in Figure 3.3.1. The basic concept and theory of adaptive lattice prediction are outlined in the following section.

3.3.1 Adaptive Lattice Prediction

The sequentially adaptive lattice predictor has the structure shown in Figure 3.3.2. s(n) is the input signal to the predictor. The aim of the predictor is to filter the signal s(n) such that the output residual f_p(n) is the minimum. f_m(n) is the forward residual, b_m(n) is the backward residual and k_m is the partial correlation (PARCOR) coefficient or reflection coefficient of the mth section of the predictor. In the lattice formulation, the PARCOR coefficients can be computed by minimizing some norm of the forward residual f_m(n) or the backward residual b_m(n), or a combination of the two (52). From Figure 3.3.2, the following relation holds:
Figure 3.3.1 System block diagram of an ADPCM coder with 6-th order adaptive lattice predictor and Jayant quantizer
Figure 3.3.2 Block diagram of an adaptive lattice predictor
\[ f_{o}(n) = b_{o}(n) = s(n) \]  
\[ f_{m+1}(n) = f_{m}(n) + K_{M+1} b_{m}(n-1) \]  
\[ b_{M+1}(n) = K_{m+1} f_{m}(n) + b_{m}(n-1) \]

where \( s(n) \) is the input signal, \( f_{p}(n) \) and \( b_{p}(n) \) are the forward and backward residual respectively.

To determine the PARCOR coefficients, we first give the following definitions:

\[ F_{m}(n) = E\left[f_{m}^{2}(n)\right] \]  
\[ B_{m}(n) = E\left[b_{m}^{2}(n)\right] \]  
\[ C_{m}(n) = E\left[f_{m}(n)b_{m}(n-1)\right] \]

where \( E\left[\cdot\right] \) denotes expected value.

A. Forward Method

In this method, the PARCOR coefficient at stage \( m+1 \) is obtained by minimizing the variance of the forward residual

\[ F_{m+1}(n) = E\left[f_{m+1}^{2}(n)\right] \]  

From equations (3.3.7) and (3.3.2) one obtains

\[ K_{m+1} f_{m} = \frac{E[f_{m}(n)b_{m}(n-1)]}{E[b_{m}^{2}(n-1)]} = \frac{C_{m}(n)}{B_{m}(n-1)} \]
B. Backward Method

In this method, the PARCOR coefficient, $K_{m+1}^b$, at stage $m+1$ is obtained by minimizing the variance of the backward residual

$$B_{m+1}(n) = E\left[b_{m+1}^2(n)\right] \quad (3.3.9)$$

From equations (3.3.9) and (3.3.3) one obtains

$$K_{m+1}^b = -\frac{E[f_m(n)b_{m}(n-1)]}{E[f_m^2(n)]} = -\frac{C_m(n)}{F_m(n)} \quad (3.3.10)$$

C. Geometric-Mean Method (Itakura)

The main problem in the previous two techniques is that the PARCOR coefficients are not always guaranteed to be less than 1 in magnitude, i.e. the stability of the inverse filter $H(z) = 1/(1-A(x))$ is not guaranteed (52). One solution to this problem was offered by Itakura (142) where the PARCOR coefficient, $K_{m+1}^I$, at stage $m+1$ is computed from

$$K_{m+1}^I = -\frac{E[f_m(n)b_{m}(n-1)]}{\sqrt{E[f_m^2(n)]E[b_m^2(n-1)]}} \quad (3.3.11)$$

$$= -\frac{C_m(n)}{\sqrt{F_m(n)B_m(n-1)}}$$
To approximate $K_{m+1}$, Itakura used

$$K_{m+1}^B = \frac{2C_m(n)}{F_m(n) + B_m(n-1)}$$  \hspace{1cm} (3.3.12)

which was also proposed by Burg\(^{(143)}\). One can easily show that $K_{m+1}^B$ can be obtained by minimizing the sum of the variances of the forward and backward residuals,

$$E_{m+1}(n) = F_{m+1}(n) + B_{m+1}(n)$$  \hspace{1cm} (3.3.13)

Burg's method is also known as the Harmonic-Mean Method.

To estimate $K_{m+1}^B$ (the subscript will be dropped hereafter), one can replace the expectation by averages and arrive at

$$K_{m+1} = -2 \frac{\sum_{n=0}^{n-1+M} f_m(n)b_m(n-1)}{\sum_{n=0}^{n-1+M} (f_m^2(n) + b_m^2(n-1))}$$ \hspace{1cm} (3.3.14)

To further simplify the computation of the equation (3.3.14), the summation terms can be approximated by the following recursive equations

$$K_{m+1}(n+1) = \frac{N_m(n)}{D_m(n)}$$ \hspace{1cm} (3.3.15)

$$N_m(n) = (1 - \gamma)N_m(n-1) - 2\gamma f_m(n)b_m(n-1)$$ \hspace{1cm} (3.3.16)

$$D_m(n) = (1 - \gamma)D_m(n-1) + \gamma[f_m^2(n) + b_m^2(n-1)]$$ \hspace{1cm} (3.3.17)
where $\gamma = 2^{-k}$ ($k = 6$ or other positive integer) is usually called the adaptation constant of the predictor.

The equations (3.3.15 - 3.3.17) represent the direct sequential method for the updating of the PARCOR coefficient.

3.3.2 The Sign Product Method

The direct sequential method presented in the previous section involves real number multiplications and divisions which are computationally expensive. They can be eliminated by invoking the very elegant relation (see Appendix I).

$$K_m(n) = \sin\left(\frac{\pi}{2} K_m^s(n)\right) \quad (3.3.18)$$

derived by J. H. Van Vleck (144). $K_m^s(n)$ is the reflection coefficient computed from the two sequences $\{f_m^s(n)\}$ and $\{b_m^s(n)\}$ which are the signs of $\{f_m(n)\}$ and $\{b_m(n)\}$ respectively.

By using Equation (3.3.14), $K_{m+1}^s(n)$ is given by

$$K_{m+1}^s(n) = -2 \frac{n^{-1+M} \sum_{n=0}^{n^{-1+M}} f_m^s(n) b_m^s(n-1)}{\sum_{n=0}^{n^{-1+M}} \left[f_m^s(n)^2 + b_m^s(n-1)^2\right]}$$

$$= - \frac{1}{n^{-1+M} \sum_{n=0}^{n^{-1+M}} f_m^s(n) b_m^s(n-1)} \quad (3.3.19)$$
Equation (3.3.19) can be approximated by the recursive equation

\[ K_{m+1}^S(n) = (1 - \gamma) K_{m+1}^S(n-1) - \gamma f_m^S(n) b_m^S(n-1) \]  \hspace{1cm} (3.3.20)

The calculation of \( K_{m+1}^S \) does not involve any real number multiplication as \( f_m^S \) and \( b_m^S \) are either equal to +1 or -1 and the multiplication by \( 2^{-K} \) (where \( K \) is integer) only requires binary shift. Finally, \( K_m^S \) can be calculated from \( b_m^S \) using Equation (3.3.18) and a sine look-up table.

In the study of ADPCM with adaptive prediction for the coding of wideband speech at 64 kbps, the 6th-order adaptive lattice predictor using the sign-product method for updating the reflection coefficients was employed. As the predictors at the encoder and decoder operate on the past decoded samples, a mean must be provided to facilitate convergence to the fixed reflection coefficients in case of transmission errors. This is necessary because transmission errors force the decoded samples at the decoder to be different from that at the encoder and consequently the reflection coefficients at the decoder may diverge from that at the encoder. A leakage factor \( \varepsilon \) can be incorporated into the calculation of \( K_m^S \) by the equation

\[ K_m^S(n) = \varepsilon K_m^S(n) + (1 - \varepsilon) K_{F_m}^S \]  \hspace{1cm} (3.3.21)

where \( K_{F_m}^S \) represent the fixed reflection coefficients, to facilitate convergence. The constant \( \varepsilon \) is usually set to \( 1 - 2^{-K} \) where \( K \) is a positive integer number usually having a value of 4 to 7.
3.4 ADPCM CODER EMPLOYING SIMPLE NOISE SPECTRAL SHAPING

The use of a pre-filter in the form of an all-zero filter with fixed coefficients and its inverse as a post-filter is perhaps the simplest means to achieve fixed spectral shaping of the output noise of an ADPCM coder. They are better known as pre-emphasis and de-emphasis filters. A simple first order pre-emphasis filter has the transfer function

\[ C(z) = 1 - a_1 z^{-1} \]  \hspace{1cm} (3.4.1)

where \( a_1 \) is a constant less than one. For the application of wideband speech, \( a_1 \) has a typical value of 0.8. The value of 0.8 arises from the fact that the corresponding de-emphasis filter, i.e. \( 1/C(z) \), has frequency response which is similar to the long-term average frequency spectrum of the speech signals.

An ADPCM coder employing fixed first-order pre-emphasis and de-emphasis to achieve noise spectral shaping is shown schematically in Figure 3.4.1. The pre-emphasis filter with \( a_1 = 0.8 \), is essentially a high-pass filter which has the effect of whitening the speech spectrum. Its frequency response is shown in Figure (3.4.2). The pre-emphasised speech is encoded by ADPCM/AQJ coder with a fixed first-order predictor (FFOP) whose coefficient is optimized with respect to the pre-emphasised speech. (Prediction coefficients of the 2nd, 3rd and 4th-order predictors optimized with respect to the pre-emphasised speech are given in Table 3.4.1.)
The decoded signal before de-emphasis consists of two components, namely the pre-emphasised signal \(x(z)c(z)\), where \(x(z)\) is the original input signal, and the quantization noise \(Q(z)\) introduced by the quantization process. If \(\hat{x}(z)\) is the input to the de-emphasis filter, then

\[\hat{x}(z) = x(z)c(z) + Q(z)\]  \hspace{1cm} (3.4.2)

As the quantization noise is relatively flat with respect to the speech spectrum, the spectral shape of the output noise of the coder is determined by the inverse filter \(1/c(z)\) because

\[R(z) = x(z) + N(z) = x(z) + Q(z)/c(z)\]  \hspace{1cm} (3.4.3)

For the ADPCM coder without pre- and de-emphasis, the recovered speech \(R(z)\) is given by

\[R(z) = x(z) + Q'(z)\]  \hspace{1cm} (3.4.4)

where \(Q'(z)\) has a greater variance than \(Q(z)\) because the pre-emphasised speech signals have a lower power and dynamic range than that of the original signals.

The spectrally shaped output noise in Equation (3.4.3) has more energy concentrating in the lower part of the spectrum (see Figure 3.4.2) and less energy in the higher part of the spectrum. As speech signals, especially voiced speech, has a similar pattern of energy distribution (see Figure 2.1.7), masking of noise by speech signals is more
Figure 3.4.1  ADPCM/AQJ with pre- and de-emphasis

Figure 3.4.2  Frequency response of the pre-emphasis and de-emphasis filters with the coefficient of 0.8
effectively achieved for the shaped noise spectrum than in the case of the flat noise spectrum. The recovered speech is therefore perceptually better than that of the coder without noise shaping.

Another effect of pre-emphasis is the reduction in the dynamic range and power of the samples to be encoded. This reduction leads to a better SNR performance across the quantizer. However, the effect of the de-emphasis filter is that of an integrator accumulating noise below the cross-over frequency of 3 kHz and attenuating noise above 3 kHz. The net effect of the dynamic range and power reduction and the accumulation of noise lead to a slight decrease in SNR performance though the subjective quality is improved.

<table>
<thead>
<tr>
<th>Predictor Order</th>
<th>Predictor Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$a_1$</td>
</tr>
<tr>
<td>1</td>
<td>0.9626</td>
</tr>
<tr>
<td>2</td>
<td>1.3961</td>
</tr>
<tr>
<td>3</td>
<td>1.3435</td>
</tr>
<tr>
<td>4</td>
<td>1.3193</td>
</tr>
</tbody>
</table>

Table 3.4.1 Predictor Coefficients for pre-emphasised wide-band speech at 16 kHz sampling rate
3.5 TWO-BAND SUBBAND CODING WITH ADPCM ENCODERS

Subband coding (SBC) differs from the traditional time-domain waveform coding technique in that the full band speech signal is first divided into a number of frequency subbands through the use of filter-bank analysis followed by the encoding of each subband signals by time-domain waveform coders. It has received considerable attentions by speech coding researchers since its first publication by Crochiere (68) in 1976. The basic factors that determine the performance of a subband coder are the number of subbands and the bandwidth of each subband, the bit allocation strategy for the quantization of the subband signals and the encoding algorithm employed for each band. Coder complexity is invariably linked to the choice of the various factors mentioned. The use of higher number of subbands, adaptive rather than fixed bit allocation and forward rather than backward adaptive quantization generally lead to higher system complexity and delay.

The relative success of subband coding over most time-domain coders lies in the fact that the system design provides the possibilities of controlling the output noise with respect to the time-changing speech spectrum so that the masking of noise by the speech signals is more effectively exploited. Furthermore, the containment of noise to each band can prevent the masking of speech signal in one frequency range by quantization noise in another frequency range.
A subband coder with higher system complexity generally has more flexibility in the control of noise and hence provides better system performance. Detailed discussions of subband coding are presented in Chapter 5.

As a relatively high number of 4 bits/sample are affordable for the design of the 64 kbps system, it is reckoned that a 2-band subband coder employing simple ADPCM coding of the subband signals is very likely to produce satisfactory subjective results. Figure 3.5.1 shows the system block diagram of a 64 kbps 2-band subband coder which employs ADPCM/AQJ coding for the upper and lower band signals. The full band signal is first divided into two equal subbands by means of Quadrature Mirror Filters (QMF). The use of QMF filter-bank analysis eliminated the requirement for sharp cut-off filters used in the early designs of subband coders. Consequently, the length of the FIR filters, which use the design concept of QMF, can be reduced to 32 with the guarantee of aliasing-free reconstruction at the receiver. The impulse response of the 32-tap prototype low-pass filter (LPF) is given in Chapter 5. The overall delay introduced by the analysis filter-bank at the transmitter and synthesis filter bank at the receiver can be easily shown to be equal to 31 samples which is equivalent to about 2 msecs.

The sampling rate of the output signals of the lower and upper bands is reduced to 8 kHz by 2:1 decimation. The lower band (0 - 4 kHz)
signal is encoded by an ADPCM coder employing fixed second order predictor (FSOP) and a 5-bit AQJ. The upper band (4-8 kHz) signal is encoded by an ADPCM coder employing fixed first order predictor (FFOP) and a 3-bit AQJ. This bit allocation is time-invariant and chosen purely on a subjective basis. At the receiver, the sampling rate is increased to 16 kHz by inserting zeros between every two decoded samples. Finally, the prediction coefficients for the two coders are given in the following Table.

<table>
<thead>
<tr>
<th>Prediction Coefficient</th>
<th>$a_1$</th>
<th>$a_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower band</td>
<td>1.3507</td>
<td>-0.4608</td>
</tr>
<tr>
<td>Upper band</td>
<td>-0.6004</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.5.1 Prediction coefficients for the 2-band SBC system

3.5.1 Two-band SBC Coder with Noise Spectral Shaping

To enhance the performance of the two-band SBC coder, the simplest noise spectral shaping of pre- and de-emphasis (PDE) is incorporated into the encoding of the lower band signal. Fixed first order pre-emphasis with the coefficient of 0.4 is used to achieve a general spectral shape for the lower band quantization noise in order to obtain a perceptually more palatable effect for the reconstructed speech.
The system block diagram of such a scheme is shown in Figure 3.5.2.

A 5-bit ADPCM/AQJ coder employing a fixed second order predictor with the prediction coefficients of 0.9116 and -0.0722 is used to code the lower band down-sampled and pre-emphasised signal. No noise shaping is applied to the upper band as quantization noise introduced by the 3-bit ADPCM coder is sufficiently low to be effectively masked by the noise-like speech signal in that band.
Figure 3.5.1 Two-band SBC with ADPCM/AQJ

Figure 3.5.2 Two-band SBC with ADPCM/AQJ and PDE
3.6 RESULTS AND DISCUSSIONS

3.6.1 ADPCM/AQJ with Fixed First/Second Order Prediction (FFOP/FSOP)

The operation of ADPCM/AQJ with fixed first/second order prediction was simulated on a PPD11/34 computer using four wideband speech sentences* as input signals. Average segmental SNR measurements were obtained for the two systems and they are listed in Table 3.6.1. The long-term average output noise spectra of the systems, together with the long-term speech spectrum are shown in Figure 3.6.1. The system with second order prediction has an average segmental SNR advantage of 2.5 dB over the system with fixed first order prediction. The output noise spectra also show a difference of about 2.5 dB difference throughout the whole frequency range (0-7 kHz). The noise spectra are relatively flat compared to the speech spectrum. This implies that there is little or no correlation between the speech signals and the output noise.

Informal subjective listening tests by a number of experienced listeners show that the distortion in the recovered speech comes mainly from the

* Sentence No. 1: Male speech "There was an old man called Michael Finnegan"
No. 2: Female speech "He grew whiskers on his chinigen"
No. 3: Male speech "She was born in 1899, in the comfortable suburb of Christchurch, New Zealand"
No. 4: Female speech "French voters have been voting in the local government elections today"
Table 3.6.1 Average segmental SNR values for ADPCM/AQJ coding of wideband speech at 64 kpbs

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>(FFOP) SNRSEG (dB)</th>
<th>(FSOP) SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.74</td>
<td>41.73</td>
</tr>
<tr>
<td>2</td>
<td>34.21</td>
<td>37.63</td>
</tr>
<tr>
<td>3</td>
<td>30.90</td>
<td>32.45</td>
</tr>
<tr>
<td>4</td>
<td>27.04</td>
<td>27.22</td>
</tr>
<tr>
<td>Average</td>
<td>32.22</td>
<td>34.76</td>
</tr>
</tbody>
</table>

Figure 3.6.1 Long-term average power spectral density plots of
(1) Original speech  
(2) Quantization noise of ADPCM/FFOP  
(3) Quantization noise of ADPCM/FSOP
high frequency end in the form of a 'hissing' noise. As the system with fixed first-order prediction (FFOP) has a higher level of noise, the 'hissing' noise is more pronounced than the system with fixed second order prediction (FSOP). As can be seen from the noise spectral plot, the average output noise of the coder with FFOP exceeds slightly the average speech power from 6.3 kHz onwards. Though the average noise spectrum for the coder with FSOP is below that of speech throughout the frequency range of 0-7 kHz, slight 'hissing' noise still can be perceived for some segments of voiced speech. This is due to the fact that the spectral plot is an average of many segments (264) of output noise spectra and therefore the short-term performance from segment to segment is not revealed.

No low frequency distortion in the form of 'rumbling' noise could be perceived for both coders due primarily to the effective masking of the noise by the speech power at the lower part of the spectrum. The near satisfactory perceptual quality of the recovered speech of the coder with FSOP leads one to believe that minimal increase in system complexity is very likely to achieve excellent quality for the coding of wideband speech at such a bit rate.

3.6.2 ADPCM/AQJ with Adaptive Lattice Prediction

Computer simulations of the ADPCM/AQJ coder with 6th-order lattice prediction using the sign product method for updating the reflection coefficients were carried out using the set of input data given in
Section 3.6.2. The leakage constants of the predictor and Jayant's quantizer and the adaptation constant of the predictor were set to the following values:

<table>
<thead>
<tr>
<th></th>
<th>Leakage constant</th>
<th>Adaptation constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictor</td>
<td>$\epsilon = 1-2^{-6}$ (Eqn. 3.3.21)</td>
<td>$\gamma = 2^{-6}$ (Eqn. 3.3.20)</td>
</tr>
<tr>
<td>Quantizer</td>
<td>$\beta = 1-2^{-8}$ (Eqn. 3.2.17)</td>
<td></td>
</tr>
</tbody>
</table>

The choice of leakage constants does not affect much the performance of the coder. More severe leakage like $\epsilon = 1-2^{-4}$ and $\beta = 1-2^{-7}$ will only reduce the SNR performance of the coder by less than 1 dB compared to the coder using the set of values given above. However, its choice is vital under transmission error conditions, as will be shown in Section 3.7.

The average segmental SNR performance of the coder for the four sentences are given in Table 3.6.2. The long-term average output noise spectrum of the system was also obtained and shown in Figure 3.6.2 for comparison with that of the coder employing fixed second order prediction. The system has more than 2 dB SNR improvement over that with FSOP. This improvement leads to a recovered speech which has an improved perceptual quality. Informal subjective listening tests show that it is very difficult to distinguish the decoded speech from the original.
Table 3.6.2  Average segmental SNR values for ADPCM/AQJ with a 6th order lattice predictor for the coding of wideband speech at 64 kbps

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42.89</td>
</tr>
<tr>
<td>2</td>
<td>39.40</td>
</tr>
<tr>
<td>3</td>
<td>34.71</td>
</tr>
<tr>
<td>4</td>
<td>30.43</td>
</tr>
<tr>
<td>Average</td>
<td>36.86</td>
</tr>
</tbody>
</table>

Figure 3.6.2  Long-term average power spectrum density plots of
(1) Original speech
(2) Output noise of the ADPCM/AQJ/Adaptive Lattice Predictor
(3) Output noise of ADPCM/AQJ/FSOP
3.6.3 ADPCM/AQJ with Simple Noise Spectral Shaping

Spectral shaping of the quantization noise of an ADPCM/AQJ/FFOP coder using simple first order pre-emphasis and de-emphasis (PDE) with the coefficient of 0.8 was simulated on a computer. The quantization noise across the quantizer of the system with PDE is lower than that of the ADPCM coder without PDE as shown in Figure 3.6.3. After passing through the de-emphasis filter, it has a spectral shape which corresponds to the frequency response of the de-emphasis filter. It can be seen that the spectral shape of the output noise now matches well with the long-term average frequency spectrum of speech.

Subjectively, the high frequency quantization 'hiss' is markedly suppressed and masked by the speech signal but at the expense of an increase in low frequency noise. The increase in low frequency 'rumbling' noise is masked by the speech power to a satisfactory degree. A very slight 'roughness' in quality could be perceived for a small number of segments of speech which has very little low frequency energy to mask the increase in low frequency noise. As a whole, it is perceptually better than the other two coders with fixed prediction. This is more so when listening through a loud-speaker which tends to reduce the effect of low frequency noise but reproduce the high frequency content of the recovered speech very faithfully. However, it is perceptually slightly inferior than that encoded by the ADPCM/AQJ...
Table 3.6.3  SNRSEG performance of the ADPCM/AQJ/FFOP/PDE coder

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38.15</td>
</tr>
<tr>
<td>2</td>
<td>34.22</td>
</tr>
<tr>
<td>3</td>
<td>29.48</td>
</tr>
<tr>
<td>4</td>
<td>23.81</td>
</tr>
<tr>
<td>Average</td>
<td>31.42</td>
</tr>
</tbody>
</table>

Figure 3.6.4  Long-term average power spectral density plots of
(1) Original speech
(2) Quantization noise of ADPCM/FFOP
(3) Quantization noise across the quantizer of ADPCM/FFOP/PDE
(4) Output noise of ADPCM/FFOP/PDE

KHZ
coder employing adaptive lattice prediction due to the very slight 'roughness' it has for some segments of the recovered speech.

The SNR measurements for the coder for the four sentences are given in Table 3.6.3. Its average performance is 0.8 dB lower than the system without PDE indicating that the noise accumulating effect of the de-emphasis filter is greater than the advantage of dynamic range and power reduction of the input samples due to the use of pre-emphasis filter.

3.6.4 Two-Band Subband Coding with ADPCM Coders

The operation of a two-band SBC scheme employing 5-bit ADPCM/AQJ/FSOP for the coding of the lower band signals and 3-bit ADPCM/AQJ/FFOP for the coding of the upper band signals, was simulated using the four speech sentences as input. To enhance the performance of the coder, noise spectral shaping in the form of fixed first-order pre/de-emphasis was incorporated into the lower band. The average segmental SNR performance of the two-band SBC systems with and without PDE is shown in Table 3.6.4. The long-term average spectral density plots of the output noise of the coders with and without PDE in comparison with that of the ADPCM/PDE system are shown in Figure 3.6.3.

The average segmental SNR measurements of SBC, SBC/PDE and ADPCM/PDE are about 31 dB with little significant difference among them. However, the subjective quality of the recovered speech of the two SBC coders
<table>
<thead>
<tr>
<th>Sentence No</th>
<th>2-Band SBC SNRSEG (dB)</th>
<th>2-Band SBC/PDE SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.48</td>
<td>35.85</td>
</tr>
<tr>
<td>2</td>
<td>34.15</td>
<td>33.56</td>
</tr>
<tr>
<td>3</td>
<td>28.50</td>
<td>27.96</td>
</tr>
<tr>
<td>4</td>
<td>25.85</td>
<td>25.16</td>
</tr>
<tr>
<td>Average</td>
<td>31.25</td>
<td>30.63</td>
</tr>
</tbody>
</table>

Table 3.6.4 SNRSEG performance for the 2-band SBC scheme with and without PDE

Figure 3.6.4 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of ADPCM/FFOP/PDE
(3) Output noise of SBC
(4) Output noise of SBC/PDE
is distinctively better than that of ADPCM/PDE. Thus subjective difference is not revealed by the spectral plots of the output noise of the coders because they only represent the long-term average measurements.

Speech is a non-stationary process and its spectral shapes for voiced and unvoiced sounds are very different with the former having more energy concentrating in the low frequency region and the latter in the high frequency region. Therefore, the non-adaptive noise spectral shaping techniques of pre- and de-emphasis to redistribute the quantization noise from the higher part of the spectrum to the lower part of the spectrum will result in a slight 'roughness' in the recovered speech for unvoiced sounds. However, this is not the case in subband coding because the coding of lower and upper band signals are independent of one another. The 5 bit allocation for the quantization of the lower band signal ensures that during unvoiced speech, the output noise in the lower band is low enough to be masked by the speech signal. Furthermore, the coding of the upper band signals does not produce low frequency 'rumbling' noise because the quantization noise introduced by the upper band encoder is confined to that band by the filtering effect of the synthesis filter. Similarly, the encoding of the lower band signals does not give rise to high frequency 'hissing' noise.

The allocation of 3 bits for the coding of the upper band signals was subjectively sufficient because the speech signal in that band
is noise like and the masking effect it provides for the quantization noise is found to be adequate. The allocation of 5 bits for the upper band was also found to be perceptually acceptable. Therefore, the choice of 5/3 bits for lower/upper band seems to be subjectively optimum for fixed bit allocation at this bit rate.

For the two SBC coders, the one which employs PDE to achieve additional noise spectral shaping provides a slight subjective improvement over the coder without PDE. The subjective quality of the recovered speech of the two-band SBC/PDE coder can be considered as very good. Its complexity is approximately double that of the ADPCM/fixed prediction coder. The performance of the SBC/PDE coder will be compared with that of the ADPCM/adaptive prediction coder under both ideal and noisy channel conditions in Section 3.7.

3.6.5 $\mu$-Law PCM at 128 kbps

For comparison purpose, $\mu$-law PCM coding of wideband speech at 128 kbps, i.e. 8 bits/sample, was simulated using the four speech sentences described in Section 3.6.1. As the bit rate is double that of all other coding systems investigated so far, it registers the highest average SNRSEG measurement of about 38.23 dB.

Despite its high SNR performance, the high frequency 'hissing' noise is still at the threshold of audibility. As the system produces flat noise spectrum regardless of the input speech spectrum, the high
frequency 'hissing' noise is not effectively masked during voiced sounds. However, it still provides slightly better perceptual results than ADPCM with fixed prediction. The quantization noise of both PCM and ADPCM/fixed prediction are not spectrally shaped. As such, the system having higher SNR performance tends to achieve better subjective results.

In the comparisons with ADPCM coder employing adaptive lattice prediction, PCM at 128 kbps was found to be subjectively inferior. As PCM has an average SNR measurement of 38.2 dB and that of ADPCM/Lattice is 36.9 dB and since both coders produce flat noise spectra, one may ask why PCM is still not as good as ADPCM subjectively. The answer lies in the fact that SNR performance of ADPCM/Lattice is sentence and speaker dependent whereas PCM gives a constant 38 dB performance for all sentences and speakers. For the first male sentence with very little unvoiced sounds, ADPCM/Lattice gives an SNRSEG measurement of 42.9 dB compared to 38.0 dB of PCM. The high-frequency 'hissing' noise of ADPCM for this sentence and speaker is thus much lower than PCM. As for the second female sentence with a lot of unvoiced sounds, ADPCM/Lattice gives SNRSEG measurement of 30.43 dB compared to 38.39 dB of PCM. However, both systems provide almost identical subjective quality for the recovered speech. This is because the relatively poor SNR performance of the ADPCM/Lattice coder during the unvoiced sounds does not necessarily lead to poor subjective
performance because unvoiced sounds are generally 'noise like' and the masking of noise by the speech signal can be more easily achieved. Thus, the high SNRSEG performance of PCM proved to be unnecessary in this case.

In the comparisons with the ADPCM/PDE and SBC schemes, PCM is slightly better than the former but inferior to the latter. The slightly higher frequency 'hissing' noise of PCM was found to be less objectionable than the low frequency 'rumbling' noise produced by PDE. However, the excellent control and shaping of the quantization noise of the SBC scheme has a clear advantage over PCM.

3.6.6 Final Remarks

From the computer simulation results of the various 64 kbps systems described so far, one can easily arrive at the following two best possible choices. For ADPCM coding of wideband speech without any form of noise spectral shaping, adaptive prediction was shown to be necessary in order to produce very good quality recovered speech. The use of the elegant relation between the cross-correlations of the backward and forward residues and that of the signs of the residues simplified the processing requirements for the updating of the PARCOR coefficients of the adaptive lattice predictor. Though ADPCM with PDE achieves a certain degree of noise spectral shaping and improves the subjective quality of the recovered speech, the overall perceptual quality is still not acceptable.
The other possible choice is subband coding with noise spectral shaping for the lower band. SBC was shown to be very effective in containing the quantization noise to each of the subbands. The different allocation of bits for the two bands according to perceptual criteria gives subjectively very good quality recovered speech.

So far, all the coding schemes are examined under transmission error free conditions. In reality, a digital transmission channel can introduce error rates as high as $10^{-3}$. Under this condition, a good speech coder should still maintain at least high intelligibility if not quality for the recovered speech. The robustness of the two chosen schemes, namely the ADPCM/Lattice and the two-band SBC coders, under transmission error conditions is further studied in the following Section.
3.7 PERFORMANCE OF THE ADPCM/LATTICE AND SBC CODERS UNDER TRANSMISSION ERROR CONDITIONS

When transmission errors occur in a digital communication channel, the noise in the recovered speech is the sum of the quantization noise introduced by the speech coder and the additional noise introduced by the effect of binary transmission errors on the coder. In the case of PCM, transmission errors only affect the samples in which they occur and their effect is non-propagating due to the non-recursive nature of PCM.

For an ADPCM coder employing a backward sequential adaptive predictor and a one-word memory Jayant quantizer, the correct recovery of the step-size of the quantizer and the reflection coefficients of the predictor at the receiver depends on the correct recovery of the transmitted bit streams. When no leakage is provided in the quantization and prediction algorithms, the effect of transmission errors can in principle perpetuate indefinitely and cause the quantizer and predictor at the receiver to operate independently of that at the transmitter. This 'mismatching' or 'capsizal' phenomenon (145) can be sufficiently overcome by introducing leakage to both the quantizer and predictor coefficient adaptation algorithms. In this case, the effect of transmission errors will decay to become negligible and the quantizer and predictor will converge back to the correct state after a short time interval.

In the study of the performance of ADPCM/Lattice and two-band SBC coders under transmission error conditions, different sets of leakage
constants for the quantizer and predictor were introduced and their final results compared. The kind of transmission error simulated was assumed to be uniformly distributed random errors which were generated by a uniform distribution random number generator. The highest bit error rate (BER) tested was $10^{-3}$ and the lowest $10^{-5}$.

The position of occurrence of the transmission errors in a speech sentence can lead to different SNR performance. Errors that happen during the silent interval or unvoiced sounds may not affect the coder performance very much. However, they have a more pronounced effect on the coder's performance when they happen during voiced sounds especially when they affect the sign or the most significant bit of a codeword. Because of the statistical nature of the effect of transmission errors, a total of twenty simulations of the coding algorithms were carried out at each bit error rate and the average SNRSEG measurement of the twenty simulations represents the coder performance at that bit error rate. For the bit error rate of $10^{-5}$, additional effort was made to ensure that the errors occurred only during voiced speech. The measurements therefore represent possibly the worst performance of the coder at this bit error rate.

3.7.1 Performance of the ADPCM/Lattice Coder Under Transmission Error Conditions

The operation of an ADPCM coder employing 6th-order adaptive lattice prediction and a one-word memory Jayant's quantizer under the transmission error conditions of $10^{-5}$, $10^{-4}$ and $10^{-3}$ was simulated. Two
sets of leakage constants given in Table 3.7.1 were used. The SNRSEG measurements of the system using the two different sets of constants are given in Table 3.7.2 and plotted against bit error rate (BER) in Figure 3.7.1. It is to be noted that the input speech materials were different from those used in the previous simulations. They are two different sentences* spoken by one male and one female speaker to give a total of 13 seconds (including silent intervals) of input signals. These sentences and speakers were chosen because they contain more unvoiced sounds and the speakers have higher pitch. The encoders therefore had more difficult input material to cope with.

It is clear from the SNRSEG measurement that the first set of leakage constants are too small to reduce the effect of transmission errors even at a very low bit error rate of $10^{-5}$. The propagation of the effect of transmission errors can cause a drop of SNRSEG by more than 15 dB at this BER. In this case, the main distortion undoubtedly comes from the effect of transmission errors rather than quantization errors. It appears therefore that the leakage of both the quantizer and predictor have to be increased if the coder is to remain as a possible candidate for the coding of wideband speech.

The second set of constants was found to improve the robustness of the coder considerably by making the coder less adaptive. Under no transmission error condition, the SNRSEG performance of the coder

* Male and Female sentences:

1) They went through the glass door into the garden
2) The first speaker was well dressed
Table 3.7.1 Leakage and adaptation constants for the ADPCM/Lattice system

<table>
<thead>
<tr>
<th></th>
<th>Adaptation Constant</th>
<th>Leakage Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>Predictor</td>
<td>$\gamma = 2^{-6}$</td>
</tr>
<tr>
<td></td>
<td>Quantizer</td>
<td>$\beta = 1 - 2^{-8}$</td>
</tr>
<tr>
<td>L2</td>
<td>Predictor</td>
<td>$\gamma = 2^{-6}$</td>
</tr>
<tr>
<td></td>
<td>Quantizer</td>
<td>$\beta = 1 - 2^{-7}$</td>
</tr>
</tbody>
</table>

Table 3.7.2 SNRSEG performance of ADPCM/lattice under transmission error conditions. (All measurements are in dB)
is about 1 dB less than that of the coder employing the first set of constants. However, no subjective difference between them could be perceived. As can be seen from Figure 3.7.1, the coder's performance is virtually independent of transmission errors below $10^{-4}$. SNRSEG only drops by about 1.6 dB at the BER of $10^{-4}$ compared to ideal channel conditions. However, the coder's performance deteriorates rapidly for BER of more than $10^{-4}$. At the BER of $10^{-3}$, it suffers an SNRSEG reduction of about 11 dB.

The choice of the second set of leakage constants only represents one of the possible sets of values the coder can employ. The ideal way to find the optimum set of constants is to carry out a real-time implementation of the coder and adjust the values of the leakage constants to achieve the optimum subjective results under both the ideal and noisy channel conditions.

3.7.2 Performance of the two-band SBC Coder Under Transmission Error Conditions

The operation of the two-band SBC/PDE coder under the transmission error rates of $10^{-5}$, $10^{-4}$ and $10^{-3}$ was simulated. The SNRSEG measurements of the coders are given in Table 3.7.2 and plotted against BER in Figure 3.7.1. The leakage of the Jayant quantizer was set to $\beta = 1-2^{-7}$.

The simulation results show that below the BER of $10^{-4}$, the effect of transmission errors is virtually negligible. The main distortion
below this BER is still attributed to quantization noise which is well contained in the two subbands and masked effectively by the speech power. Subjectively, hardly any distortion like coder overload in the form of "clicking" noise due to the effect of transmission error could be perceived below this BER.

When the BER was above \(10^{-6}\), the effect of transmission error was more pronounced. The coder suffers a drop of 7 dB in SNRSEG performance when the BER is increased to \(10^{-3}\). However, compared to the ADPCM/Lattice system, the slope of the SNR curve of the SBC coder from the BER of \(10^{-4}\) to \(10^{-3}\) is more gentle (Figure 3.7.1), indicating the superiority of SBC over ADPCM in the presence of transmission errors.

<table>
<thead>
<tr>
<th>BER</th>
<th>0</th>
<th>(10^{-5})</th>
<th>(10^{-4})</th>
<th>(10^{-3})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>22.53</td>
<td>22.47</td>
<td>21.50</td>
<td>16.80</td>
</tr>
<tr>
<td>Female</td>
<td>27.44</td>
<td>27.34</td>
<td>26.43</td>
<td>18.76</td>
</tr>
<tr>
<td>Average</td>
<td>24.99</td>
<td>24.91</td>
<td>23.73</td>
<td>17.78</td>
</tr>
</tbody>
</table>

*Table 3.7.2* SNRSEG performance of the 2-band SBC/PDE coder under transmission error conditions
Fig. 3.7.1 SNRSEG performance against bit error rate
3.8 REAL-TIME IMPLEMENTATION OF A TWO-BAND SUBBAND CODER

Until recently, evaluation of digital processing techniques was limited to non real-time computer simulations. With the rapid advancements in digital hardware and the advent of microprocessor and Digital Signal Processing (DSP) devices, the real-time implementation of relatively sophisticated speech coding algorithms became feasible. The two-band SBC coder described in Section 3.5 was realized in real-time by D. Lee of British Telecom and S.J. Perkins of Loughborough University (146-147) using a single NEC 7720 DSP chip. Their implementation provides an indication of the performance of the SBC coder operating in real-time when constrained by the various limitations imposed by the currently available DSP technology.

The NEC UPD 7720 is a single chip microprocessor optimized for signal processing algorithms. An architectural block diagram of the device is shown in Figure 3.8.1. It is a 16-bit device with a 16-bit Arithmetic Logic Unit (ALU) and a separate 16 x 16 bit multiplier. On board memory is divided into programme ROM, data ROM and data RAM. The 512 x 23 bits of programme ROM, are addressed by a 9-bit programme counter which can be modified by external reset, interrupt, call or return instructions. The data ROM is organised into 512 x 13 bit words and is also addressed through a 9-bit ROM pointer (RP) which may be modified as part of an arithmetic instruction. The data RAM is
Figure 3.8.1 Architectural block diagram of the 7720 DSP Chip
organised into 128 x 16 bit words and addressed through a data pointer (DP) which may also be modified by the arithmetic instructions. 

The 7720 instruction sets can be divided into three types. They are the arithmetic, jump/call and load instructions. One major drawback of this device is that the operation of division cannot be implemented easily. An indirect approach is via multiplication, or binary shift if the divisor is a multiple of two, to achieve division operations.

In the non real-time computer simulations of the two-band SBC coder, the step-size of the Jayant's quantizer can practically assume any value. In the DSP chip realization, the following modification is usually made. The one-word memory AQJ algorithm given by

$$\Delta_{i+1} = \Delta_i \beta M(C_i) \quad (3.8.1)$$

is implemented in its logarithmic form:

$$\log_e \Delta_{i+1} = \beta \log_e \Delta_i + \log_e M(C_i) \quad (3.8.2)$$

where $\Delta_i$ is the step-size at the $i$th sample. $\beta$ is the leakage factor, and $M(C_i)$ is the multiplier function of the previous codeword $C_i$.

To calculate the actual step-size, it is necessary to store the exponential function in the ROM look-up tables in the DSP chip.

The finite (512) ROM memory locations used for the calculation of the step-size implies that the step-size of the quantizer is restricted to
the maximum possible of 512 values contained in the table. Despite this major limitation, the quality of the recovered speech of the SBC coder operating in real-time was found to be as good as that from the non real-time computer simulations. This indicates the potential of employing the SBC technique, implemented using the present DSP technology, for high quality enhanced bandwidth speech communication for a variety of applications.
3.9 PERFORMANCE OF A TWO-BAND SUBBAND CODER OPERATING AT 56 kbps

Non real-time computer simulations and real-time DSP implementation of the two-band SBC/PDE system show that it is feasible to obtain very good quality wideband speech at 64 kbps using relatively simple time or frequency-domain waveform coding techniques. Though one of the primary provisions in the forthcoming ISDN is the transmission of wideband speech at 64 kbps, there might arise the need to transmit 8 or 16 kbps of data together with wideband speech within the same 64 kbps channel. This additional requirement implies that the bit rate available for the transmission of wideband speech is reduced to 56 or 48 kbps.

In the case of two band SBC coder, one way to transmit wideband speech at 56 or 48 kbps while maintaining the quality of the recovered speech as that at 64 kbps is to employ adaptive bit allocation (Chapter 5) instead of fixing the number of bits allocated to the lower and upper bands to be 5 and 3 respectively. Alternatively, the efficiency of the lower band encoder can be improved by employing adaptive prediction like the adaptive lattice prediction described in Section 3.3. The number of bits allocated to the lower band can then be reduced to 4 and thus giving a total of 56 kbps for the system. Fixed bit allocation has the advantage that no additional side-information is required, compared to the coder which employs 'forward' adaptive bit allocation.
Though 'backward' adaptive strategies can be used to avoid the need to transmit side-information, coders employing this type of adaptation is less robust than those employing 'forward' adaptation in the presence of transmission errors.

A two-band subband coder employing an ADPCM coder with 6th-order adaptive lattice prediction and a 4-bit AQJ for the coding of the lower band signals and another ADPCM coder with fixed first order prediction and a 3-bit AQJ for the coding of the upper band signals was simulated on a computer. The sign-product method was employed for the updating of the reflection coefficients of the lattice predictor. Noise shaping in the form of pre-emphasis and de-emphasis was also incorporated into the lower band. The leakage and adaptation constants of the predictor and the quantizer were set to that of L2 given in Table 3.7.1. The SNRSEG measurements of the coder under ideal channel conditions was found to be an average of 22.41 dB. To assess its performance under noisy channel conditions, bit error rates of $10^{-5}$, $10^{-4}$ and $10^{-3}$ were introduced in the simulations. Table 3.9.1 gives a summary of the SNRSEG measurements of the system for the various BERs. Its performance in comparison with the two-band SBC/PDE and ADPCM/Lattice is shown in Figure 3.9.1.

Under no transmission error condition, the 56 kbps system is lower than the 64 kbps SBC/PDE scheme by about 2.6 dB in SNRSEG measurements. This difference decreases to less than one dB at the BER of $10^{-3}$. Subjectively,
Table 3.9.1 SNRSEG measurements for the 56 kbps 2-band subband coder under transmission error conditions

<table>
<thead>
<tr>
<th>BER</th>
<th>0</th>
<th>$10^{-5}$</th>
<th>$10^{-4}$</th>
<th>$10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>19.72</td>
<td>19.69</td>
<td>10.10</td>
<td>15.98</td>
</tr>
<tr>
<td>Female</td>
<td>25.10</td>
<td>25.02</td>
<td>24.24</td>
<td>16.89</td>
</tr>
<tr>
<td>Average</td>
<td>22.41</td>
<td>22.36</td>
<td>21.67</td>
<td>16.94</td>
</tr>
</tbody>
</table>

Figure 3.9.1 SNRSEG performance of the two 64 kbps systems (L2 and SBC/PDE) and the 56 kbps system (SBC56) against bit error rates
the quality of the recovered speech of the coder was found to be very close to that of the 64 kbps schemes with only a very slight increase in the audibility of quantization noise in the lower band. The effect of transmission error at or below $10^{-4}$ was found to be negligible. At the BER of $10^{-3}$, the recovered speech was found to be as annoying as the other schemes operating at 64 kbps. At this BER, the effect of transmission errors outweighs that of the quantization noise.

3.10 WIDEBAND SPEECH CODING RESEARCH WITHIN THE CCITT

Due to the importance of wideband speech applications in the forthcoming Integrated Services Digital Networks, a rapporteur's group consisting of the representatives from the various telecommunication administrations has been set up recently within CCITT\(^{(120-123)}\) to study and recommend the various 64 kbps coding algorithms for the transmission of wideband speech. Indeed, the ADPCM/PDE and 2-band SBC/PDE systems described so far in this chapter formed the initial contributions of British Telecom to the study group\(^{(150)}\). The ADPCM coder with adaptive lattice prediction, described in Section 3.3, was proposed by the French P.T.T. as a possible candidate\(^{(148)}\). In contrast to the relatively simple 2-band SBC and ADPCM proposals, the Italian P.T.T.\(^{(149)}\) suggested the use of Time Domain Harmonic Scaling to reduce the bandwidth of the input speech by half. The compressed signal is then downsampled to 8 kHz and quantized by 8 bit PCM to achieve a transmission bit rate of 64 kbps.
More recently, the objective of the group has been extended to study speech coding algorithms which will allow 8 or 16 kbps of data to be multiplexed with the digitized speech, producing a total bit rate of 64 kbps (120-123). Thus the speech bit rate is reduced to 56 or 48 kbps. In response to this new requirement, the following new proposals and modifications to the original algorithms were suggested by the group members:

1. B.T. proposed the use of a very simple 'backward' adaptive bit allocation algorithm for the 2-band SBC coder. To operate at 56 kbps, the coder allocates either 5 and 2 or 4 and 3 bits instead of the fixed allocation of 5 and 3 bits to the lower and upper-band respectively, depending on the scaled ratio of the previous lower and upper-band adaptive Jayant's quantizer step-sizes (120). The aim is to embed one bit of data to either of the bands depending on whether voiced or unvoiced speech is present. Fixed prediction is still retained for both the lower and upper-band ADPCM coders.

2. Instead of using fixed prediction, the French P.T.T. proposed the use of adaptive pole-zero prediction for the ADPCM coder in both the lower and upper-band of a 2-band SBC coder. The bit allocation on the other hand, is fixed to 5 and 3 for 64 kbps and 5 and 2 for 56 kbps. In addition to that, one out of the $2^5$ code-words is deleted from the lower-band and used as a 'flap' to inform the receiver about the switching of the speech transmission bit rate (122).
Similar to the French proposal, N.T.T. of Japan suggested the fixed bit allocation of 6 and 2, 5 and 2 or 4 and 2 for the lower and upper-band when operating at 64, 56 and 48 kbps respectively. The ADPCM coders in both bands employ adaptive pole-zero predictions (123).

The algorithms outlined above have yet to be finalized and rigorously tested, as the exact system requirements, i.e. whether to operate with two or three different bit rates, has not been decided to date. Nevertheless, a few comments can be made about the present proposals. The design strategy of B.T. is to exploit the time-varying characteristics of speech through the use of adaptive bit allocation. However, as the bit allocation algorithm operates in a 'backward' mode, transmission errors, especially burst errors, may force the receiver to operate with different bit allocation patterns than those at the transmitter. Therefore, the robustness of the bit allocation algorithm under noisy channel conditions has to be thoroughly investigated.

The French strategy is to employ adaptive prediction for the ADPCM coders in both bands in order to adapt to the time-varying characteristics of the subband signals. To operate at 56 kbps, 2 bits are allocated to the upper-band at all times. Though 2 bits are sufficient during voiced speech which has little energy in the upper-band, the quality of unvoiced speech with considerable energy in the upper-band, is reduced. Similar comment is also applicable to the N.T.T. proposal.
While more modifications and possible improvements to the present algorithms are expected to be proposed by various administrations in the near future, the rest of this thesis presents a study of the various time and frequency-domain coding techniques operating at the even lower bit rate of 32 kbps. The ultimate aim is to reduce the bit rate to 32 kbps while maintaining the speech quality equal to that obtained by the 64 kbps coders.

3.11 NOTE ON PUBLICATIONS

(1) A paper entitled, "64 kbit/s coding of 7 kHz bandwidth speech" was presented at the colloquium on "Digital Processing of Speech" organised by IEE and held in London. It was written in co-authorship with Dr C. S. Xydeas, Mr S. Perkin of Loughborough University and Mr D. Lee of British Telecom. The paper summarizes the work described in this chapter.

(2) A second paper, similar to the first one, entitled "Real-time implementation of 64 kbits/sec wideband speech coders using the NEC 7720 processor" was presented at the 1983 KRIKOS-OTE Telecommunication Conference held in Athens, Greece. The paper summarises the work described in this chapter with the inclusion of the performance of the coders under noisy channel conditions. It was recorded in the Conference Proceedings.
This chapter has provided a comparative study of the performance of the various simple time and frequency-domain waveform coding techniques for the coding of wideband speech at 64 and 56 kbps under both the ideal and transmission error conditions. At 64 kbps, the recovered speech of the ADPCM coder employing either the fixed first order or fixed second order predictor was found to be slightly degraded by a small amount of 'hissing' noise. To improve the quality of the recovered speech of the ADPCM coder, two different approaches, namely noise spectral shaping via pre- and de-emphasis and adaptive lattice prediction, were examined. The use of pre- and de-emphasis very successively cuts down the high frequency 'hissing' noise but at the expense of an increase in low frequency noise which is partially masked by the speech spectrum. An overall subjective improvement is evident when using PDE though there is a drop in SNR performance. The second approach of employing adaptive lattice prediction leads to improvements in both SNR and perceptual performance. The recovered speech was found to be very difficult to be differentiated from the original unprocessed speech.

In the frequency-domain waveform category, two-band subband coding employing fixed bit allocation of 5 and 3 bits for the lower and upper band and ADPCM encoders for both bands also produces recovered speech of near excellent quality. To enhance the quality of the two-band
coder, pre- and de-emphasis was incorporated into the lower band to
further shape the output noise spectrum in that band. When compared
with ADPCM/Lattice over loudspeaker, the two band SBC coder was
marginally better. This shows that very good quality wideband speech
can be easily obtained using simple two-band subband coding or ADPCM/
lattice encoder at 64 kbps. In fact, the performance of both schemes
was shown to be even better than log PCM operating at twice the bit
rate of 128 kbps. The real-time implementation of the two-band SBC
coder using a single NEC 7720 DSP device further demonstrates that the
two-band SBC scheme is truly a viable candidate for coding wideband
speech at such a bit rate.

The robustness of both the two-band SBC and ADPCM/lattice encoder
under transmission error conditions was also examined. With proper
choice of leakage parameters for both the AQJ and adaptive lattice
predictor, the ADPCM coder was shown to suffer a drop of 11 dB in SNR,
compared to 7 dB in the case of 2-band SBC coder, at the BER of $10^{-3}$.

By replacing fixed prediction by adaptive lattice prediction in the
lower band ADPCM encoder in a two-band SBC scheme, and by reducing the
allocation of 5 bits to that band to 4, a two-band SBC coder operating
at 56 kbps is obtained. Hence 8 kbps of channel capacity can now
be employed for transmitting data. The recovered speech of the
56 kbps two-band SBC scheme is only slightly degraded by a higher
level of quantization noise in the lower band. Of course better
performance can be obtained by applying adaptive bit allocation strategy to the two-band SBC coder. However, it was decided that efforts should be channelled to the investigation of wideband speech at the even lower bit rate of 32 kbps so that the 64 kbps digital link can accommodate two instead of one wideband speech channels. In the following chapters, ADPCM with adaptive noise spectral shaping and frequency-domain coders like the more complicated forms of SBC and ATC coders will be examined for the coding of wideband speech at 32 kbps. The ultimate aim is to reduce the bit rate from 64 to 32 kbps while maintaining the quality of the recovered speech as that produced by the 64 kbps schemes.
CHAPTER 4

ADPCM CODING OF SPEECH WITH ADAPTIVE NOISE

SPECTRAL SHAPING
4.1 INTRODUCTION

In the previous chapter, simple time and frequency-domain waveform coding techniques like ADPCM with adaptive lattice prediction and two-band subband coder have been shown to be capable of producing very good quality wideband speech at 64 kbps. With a slight increase in the complexity of the two-band SBC coder, the transmission bit rate can be lowered to 56 kbps while maintaining almost the same subjective quality as that of the 64 kbps system. Clearly, coder complexity is expected to be increased considerably if the transmission bit rate is to be reduced further to 32 kbps. As there are only an average of 2 bits available for the coding of each sample, more sophisticated and efficient coding techniques are required in order to adapt to the time-varying short-term frequency characteristics of the speech signals and to exploit the auditory masking properties of the ear.

In the area of time-domain waveform coding, the technique of adaptive noise spectral shaping (ANS) can be employed to improve the perceptual quality of the recovered speech. In the frequency-domain waveform coding category, more sophisticated subband coding systems employing a large number of subbands, adaptive bit allocation and pre- and post-filtering are required to maintain the same quality while reducing the bit rate to 32 kbps. In this chapter, the effectiveness of the various combinations of noise spectral shaping techniques like pre- and de-emphasis, ANS via adaptive feedback filtering and adaptive pre- and post-filtering are examined. The differential coding scheme
that incorporates the ANS algorithms is ADPCM employing a 2-bit AQF and a 4th-order forward block adaptive predictor. The various ANS techniques are therefore examined under coarse quantization conditions as only 2 bits are available for the quantization of each sample. The subject of adaptive noise spectral shaping is presented first followed by description of the various systems examined. The performance of all the systems is assessed and compared in terms of average segmental SNR measurements, long-term average spectral density plots of the output noise and informal subjective listening tests.
4.2 ADAPTIVE PREDICTIVE CODING (APC) OF SPEECH

The adaptive lattice predictor described in the 64 kbps system seeks to remove the redundant structure in the speech signals by tracking the short-term speech spectral envelope which is mainly determined for voiced speech by the frequency response of the vocal tract. However, a spectral envelope predictor like the adaptive lattice predictor only removes part of the redundant information in the speech signals. The difference signal after prediction based on the spectral envelope of speech is still periodic in structure (Figure 4.2.1) during voiced speech. This is because voiced speech is produced by exciting the vocal tract by a train of periodic pulses of a certain pitch period and the resulting quasi-periodic structure of voiced speech gives rise to a spectral fine structure which cannot be removed by a spectral envelope predictor (134).

It is well known that to achieve the minimum quantization noise power in a differential coding scheme, the variance of the difference signal to the quantizer should be as small as possible. If the redundant periodic structure of the difference signal can be removed, the SNR performance of the coding system will certainly be improved. Adaptive Predictive Coding (APC) of speech, proposed by Atal and Schroeder (130), employs both spectral envelope prediction and pitch prediction. Two equivalent system configurations of APC are shown in Figure 4.2.2a and b. The combined predictor of APC can be expressed in the z-transform notation as

\[ A(z) = A_s(z) + A_d(z) \left[ 1 - A_s(z) \right] \]  

(4.2.1)
Figure 4.2.1  (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)  
(after Reference 134)
Figure 4.2.2 Two equivalent APC Configurations
where $A_s(z)$ and $A_d(z)$ represent the spectral envelope predictor and the pitch predictor respectively. An $N$-order spectral envelope predictor is given by

$$A_s(z) = \sum_{i=1}^{N} a_i z^{-i} \quad (4.2.2)$$

and a pitch predictor is usually implemented as

$$A_d(z) = \beta_1 z^{-M+1} + \beta_2 z^{-M} + \beta_3 z^{-M-1} \quad (4.2.3)$$

where $M$ corresponds to the pitch period and the coefficients $\beta_1$, $\beta_2$ and $\beta_3$ are determined by minimizing the mean squared difference between $d_n$ (the difference signal after the spectral envelope prediction) and its predicted value.

The relative ordering of the two kinds of prediction is important. If the positions of the envelope predictor and pitch predictor in Figure 4.2.2 are interchanged, i.e. pitch prediction is carried out first before envelope prediction, the combined predictor is given by

$$A(z) = A_d(z) + A_s(z) \left[ 1 - A_d(z) \right] \quad (4.2.4)$$

Though Equation (4.2.4) can be obtained by re-arranging the terms on the right-hand side of Equation (4.2.1) the performance of the two combined predictors given by these two equations is not exactly the same because in the first case (Equation (4.2.1)), the pitch predictor coefficients are optimized after the spectral envelope prediction whereas in the second case, the envelope predictor coefficients are optimized after the pitch prediction. Recent studies on APC by Atal (134) suggest that it is preferable to apply spectral envelope prediction before pitch prediction.
APC achieves considerably better SNR performance than ADPCM. However, it requires a pitch predictor, a delay of at least two pitch periods (about 0.2 sec or more) and the transmission of the pitch information to the receiver. It is therefore more difficult than ADPCM to be implemented in real-time. Furthermore, the reduction of noise power does not necessarily guarantee an improvement in speech quality because the short-term spectral shape of the output noise spectrum in relation to that of the speech spectrum is another key factor in determining the perceptual quality of the decoded speech. More recent studies (131-139) have been focused on the aspect of spectral shaping of the output noise of a coder via i) adaptive feedback filtering or ii) adaptive pre-filtering, to improve the perceptual quality of the recovered speech. Each of the ANS techniques can be implemented in a forward or backward mode depending on whether the adaptation is derived from the input signals or the locally decoded signals. In the forward ANS techniques, the coefficients of either the feedback filter or pre-filter have to be quantized and transmitted to the receiver as side-information. Additional bit rate is thus incurred. However, they are generally better than the backward ANS techniques. The various ANS techniques are described in the following section.
4.3 ADAPTIVE NOISE SPECTRAL SHAPING (ANS) TECHNIQUES

Differential adaptive predictive coders attempt to minimize the power of the quantization noise in the coded signal. The quantization noise bears no correlation with the speech signals and it has a relatively flat spectrum compared to the speech spectrum regardless of the shape of the latter, if there is a sufficient number of bits for the quantization of the error signals and if the quantizer is optimally designed. However, the human ear does not perceive signal distortion on the basis of rms error alone. In designing a perceptually optimum speech coder, it is necessary to consider the short-term spectrum of the output noise in relation to that of the input speech. The theory of auditory masking suggests that noise in the formant regions can be partially or totally masked by the speech signals. The perceived noise in the coder comes mainly from the region where the signal level is low. During voiced speech for instance, the speech energy in the high frequency region is usually not high enough to mask the high frequency quantization 'hissing' noise which becomes the main source of degradation. However, the low frequency 'rumbling' noise is partially or totally masked by the high concentration of speech energy in the lower part of the spectrum. This observation leads to a very sensible suggestion of shaping the otherwise relatively flat noise spectrum by re-distributing the noise energy from the frequency region where there is little speech energy to the formant regions where most speech energy is situated. Masking of noise can then be more effectively
exploited so that the decoded speech has a subjectively more pleasing quality. Figure 4.3.1 illustrates the unshaped and shaped output noise spectra in relation to the speech spectrum.

4.3.1 Adaptive Noise Spectral Shaping by Atal and Schroeder

One way of achieving noise spectral shaping is the technique of adaptive feedback filtering of the quantization noise suggested by Atal and Schroeder (131). Figure 4.3.2 shows the ADPCM/ANS configuration proposed by Atal. The quantization noise is fed back via an adaptive filter \( F(z) \) which can be expressed as

\[
F(z) = A(z/\beta)
\]

\[
= a_1 \beta z^{-1} + a_2 \beta^2 z^{-2} + a_3 \beta^3 z^{-3} + \ldots + a_N \beta^N z^{-N}
\]

\[
= b_1 z^{-1} + b_2 z^{-2} + b_3 z^{-3} + \ldots + b_N z^{-N}
\]  

(4.3.1)

where \( A(z) \) is the spectral envelope predictor with the coefficients \( a_1, a_2, \ldots, a_N \), \( N \) is the order of both the predictor and the noise feedback filter, and the noise shaping factor \( \beta \) has a value between 0 and 1. The coefficients of \( A(z) \) and \( F(z) \) are updated every short segment of speech of about 20 msec during which the speech signal is generally stationary. The choice of this particular form of \( F(z) \) would become clear in the following discussions.

The quantizer input \( e_n \) in Figure 4.3.2 can be written as

\[
e_n = s_n - \sum_{i=1}^{N} s_{n-i} a_i - \sum_{i=1}^{N} q_{n-i} b_i
\]

(4.3.2)
Figure 4.3.1 An example showing the output noise spectrum shaped to reduce perceptual distortion
Figure 4.3.2  The ADPCM/ANS configuration proposed by Atal

Figure 4.3.3  Two equivalent ADPCM configurations
where \( s_n \) is the \( n \)th input sample and \( q_n \) is the \( n \)th sample of the
quantization error which is also the input signal to the feedback
filter \( F \). The coder output \( \hat{s}_n \) is given by

\[
\hat{s}_n = e_n + q_n + \sum_{i=1}^{N} s_{n-i} a_i
\]

\[
= s_n + \sum_{i=1}^{N} (s_{n-i} a_i - s_{n-i} a_i) + (q_n - \sum_{i=1}^{N} q_{n-i} b_i) \tag{4.3.3}
\]

Representing Fourier transforms by upper case letter, Equation (4.3.3)
can be written in frequency-domain notation as

\[
\hat{S} - S = Q \frac{1-F}{1-A} \tag{4.3.4}
\]

The output noise power spectral density \( S_n(f) \) of the system is then
given by

\[
S_n(f) = S_q(f) \left[ \frac{1-F(f)}{1-A(f)} \right]^2 \tag{4.3.5}
\]

where \( S_q(f) \) is the power spectral density of the quantization noise.

The ADPCM/ANS configuration shown in Figure 4.3.2 reduces to various
interesting systems when \( F(z) \) assumes different transfer functions.

1. When \( F(z) = A(z) \), i.e. \( \beta = 1 \), the system can be shown to be
equivalent to the conventional ADPCM without any noise spectral
shaping function. The output noise spectral density of the system \( s_n(f) \) in this case is equal to the quantization noise \( s_q(f) \) which is spectrally flat if the quantizer is optimally designed.

\[(2)\] When \( F(z) = 0 \), i.e. \( \beta = 0 \), there is no feedback of the quantization noise. The output noise power of the system is then given by

\[
s_n(f) = s_q(f) \left[ \frac{1}{1-A(f)} \right]^2 \tag{4.3.6}
\]

It therefore has a similar spectral shape to that of the speech spectrum. This particular configuration is termed D*PCM by P. No\( \text{ll} \)\textsuperscript{(151)}. It is well known that D*PCM is inferior to DPCM in terms of SNR performance.

\[(3)\] When the choice of \( F(z) \) is between these two extremes of flat noise spectrum and 'completely' shaped noise spectrum, i.e. \( F(z) = A(z/\beta) \) where \( 0 < \beta < 1 \), the value of \( \beta \) can be adjusted to achieve any desired degree of noise spectral shaping for optimum perceptual effect. As \( \beta = 1 \) corresponds to no shaping and \( \beta = 0 \) corresponds to 'complete' shaping, smaller value of \( \beta \) means a higher degree shaping. Figure 4.3.4 illustrates the relation between the output noise spectrum and the speech spectrum for different values of \( \beta \).
Figure 4.3.4 Illustrations of the speech to noise spectral relation for different values of $\beta$
When the quantizer is optimally designed, the quantization noise has a flat spectrum and the spectrum of the coder output noise is determined only by the factor \((1-F)/(1-A)\) as shown in Equation (4.3.5). Let the squared magnitude of this factor at a frequency \(f\) be \(\Gamma(f)\), i.e.

\[
\Gamma(f) = \left| \frac{1 - F(e^{j2\pi fT})}{1 - A(e^{j2\pi fT})} \right|^2
\]

(4.3.7)

where \(T\) is the sampling interval. It can be shown (see Appendix II) that the following important constraint holds

\[
\frac{1}{f_s} \int_{0}^{f_s} \log \Gamma(f) \, df = 0
\]

(4.3.8)

where \(f_s\) is the sampling frequency. It means the average value of \(\log \Gamma(f)\) is zero and therefore

\[
\int_{0}^{f_s} 10 \log_{10} S_n(e^{j2\pi fT}) = \int_{0}^{f_s} 10 \log_{10} S_q(e^{j2\pi fT})
\]

(4.3.9)

The average value of the Log power spectrum of the output noise is then determined by the quantizer and is not altered by the choice of the filter \(F\) or the predictor \(A\). The filter \(F\), however, redistributes the noise power from one frequency to another. Regardless of how the noise power is re-distributed, i.e. whether the noise power is re-distributed from high frequency region to low frequency region or vice versa, the logarithmic spectrum of the quantization noise has equal area above and below the average power level determined by \(S_q(f)\). As
it is the logarithmic areas that are equal, the actual noise power is higher when there is a deviation from the flat-spectrum minimum rms noise case. Therefore, this form of noise spectral shaping will invariably lead to a decrease in the coder's SNR performance compared to the similar system without any noise shaping.

Besides the adaptive noise feedback filtering described so far, Atal and Schroeder also added fixed first-order pre-emphasis and de-emphasis filters at the beginning and at the end of the system respectively to further improve the subjective quality of the recovered speech. In this case, the final output noise spectrum is further shaped by the de-emphasis filter, i.e.

\[
S_n(f) = S_q(f) \left( \frac{1 - F(f)}{1 - A(f)} \right)^2 \left( \frac{1}{C(f)} \right)^2
\]  

(4.3.10)

where \(1/C(f)\) is the Fourier transform of the impulse response of the de-emphasis filter. The final spectral shape of the output noise of the system is now determined both by the choices of \(F\) and \(C\).

Finally, it should be noted that the quantization noise \(S_q(f)\) in Equation (4.3.10) is smaller than that in Equation (4.3.5) due to the combined prediction gain of the pre-emphasis filter \(C(z)\) and the predictor \(A(z)\). If a pitch predictor is included in the system as shown in Figure 4.3.2, an even smaller value of \(S_q(f)\) can be achieved.
4.3.2 Adaptive Noise Spectral Shaping by Makhoul and Berouti

Concurrent with the work done by Atal and Schroeder, Makhoul and Berouti also came out with a proposal for the implementation of spectral noise shaping in an ADPCM structure. From the DPCM structure in Figure (4.3.5), the following equations can be easily obtained.

\[
R(Z) = \frac{W(Z)}{1-A(Z)} \quad (4.3.11)
\]
\[
W(Z) = S(Z) - A(Z)W(Z) / (1-A(Z)) \quad (4.3.12)
\]
\[
Q(Z) = W(Z) - W(Z) \quad (4.3.13)
\]
\[
W(Z) = S(Z) (1-A(Z)) - A(Z)Q(Z) \quad (4.3.14)
\]

Using these equations, one can show that the reconstructed signal at the receiver is given by

\[
R(Z) = S(Z) + Q(Z) \quad (4.3.15)
\]

which is the sum of the input signal and the quantization noise.

To derive the DPCM system with noise shaping properties, Makhoul began by assuming that the output speech is given by

\[
R(Z) = S(Z) + Q(Z)B(Z) \quad (4.3.16)
\]

where B(Z) is the desired spectral shape of the output noise of the system. When \( R(Z) \) is defined by equation (4.3.16), the difference signal \( W(Z) \) can be shown using equations (4.3.11) and (4.3.13) to be given by

\[
W(Z) = S(Z) [1-A(Z)] - (1-B(Z) [1-A(Z)]) Q(Z) \quad (4.3.17)
\]

Equation (4.3.17) can be implemented as shown in Figure 4.3.6.
Figure 4.3.5 Conventional ADPCM/AQF configuration

Figure 4.3.6 An initial configuration for ADPCM/ANS system

Figure 4.3.7 The ADPCM/ANS configuration proposed by Makhoul
Next, by adding and subtracting the term $A(Z)B(Z)Q(Z)$ from the right hand side of equation (4.3.17), he obtained the final system equation

$$W(Z) = S(Z) - A(Z) [S(Z) + B(Z)Q(Z)] - Q(Z)[1-B(Z)]$$

(4.3.18)

The final system configuration is shown in Figure 4.3.7. One important property of this implementation is the decoupling of $A(Z)$ and $B(Z)$. It is also important to note that the noise shaping filter only exists in the encoder and there is no corresponding "inverse filter" at the receiver.

The remaining issue is the determination of the choice of $B(Z)$. Makhoul had experimented with different filter shapes of the form

$$B(Z) = \frac{1-A(Z/\beta)}{1-A(Z/\alpha)}$$

(4.3.19)

where $\alpha$ and $\beta$ are constants to be optimized. He found that $\alpha = 1$ and $\beta = 0.5$ gave the best perceptual results for speech sampled at 6.67 kHz and coded at 16 kbps.
4.3.3 Comparison between the ANS Techniques Proposed by Atal and Makhoul

The differences between the two adaptive noise shaping systems proposed by Atal and Schroeder and Makhoul and Berouti are

(i) The pre-emphasis and de-emphasis filters are not used in the structure proposed by Makhoul and Berouti.

(ii) The noise shaping filter is outside the DPCM feedback loop in Atal's structure as opposed to the structure which has the noise shaping filter inside the feedback loop proposed by Makhoul.

If Makhoul noise shaping technique is implemented as shown in Figure 4.3.6, the noise feedback filter $F(z)$ is given by

$$F(z) = 1 - B(z)\left[1 - A(z)\right]$$

$$= 1 - \frac{1 - A(z/\beta)}{1 - A(z)} \left[1 - A(z)\right]$$

$$= A(z/\beta)$$

which is exactly the noise shaping filter used by Atal and Schroeder.

Without the use of pre-emphasis and de-emphasis, both the Atal and Makhoul configurations achieve exactly the same noise shaping function.
4.3.4 Adaptive Noise Spectral Shaping via the Forward Adaptive Pre- and Post-filtering Technique

The third method of providing flexibility in controlling the spectral shape of the output noise of an ADPCM or APC system is to incorporate pre- and post-filtering. Such a system is shown in Figure 4.3.8. The input signal \( s_n \) is pre-filtered by a time-varying filter \( 1-R(z) \) to generate a residual signal \( y_n \). The predictor \( A(z) \) in the ADPCM or APC coder is optimized for predicting the signal \( y_n \). \( A(z) \) is simply a spectral envelope predictor \( A_s(z) \) for an ADPCM system but includes an additional pitch predictor \( A_d(z) \) in the case of APC. The noise shaping properties of this configuration can be easily seen by expressing the output noise power \( S_n(f) \) in terms of the system parameter by the equation

\[
S_n(f) = S_q(f) \left( \frac{1}{1-R(f)} \right)^2 \tag{4.3.21}
\]

where \( 1-R(f) \) is the Fourier transform of the pre-filter's impulse response and \( S_q(f) \) is the quantization noise power. It shows that any desired spectrum of the output noise can be obtained by the appropriate selection of the filter \( R(z) \) since \( S_q(f) \) is a relatively flat spectrum.

Consider the choice \( R = 0 \), i.e. no pre- and post-filtering, the system reduces to the conventional ADPCM or APC system without noise spectral shaping and \( S_n(f) \) is equal to the spectrally flat \( S_q(f) \). If \( R = A_s \),
the system achieves 'complete' noise shaping. These two choices of \( R \) correspond to \( S = 1 \) and \( S = 0 \) respectively if \( R \) is given by

\[
1 - R(z) = \frac{1 - A(z)}{1 - A(z/B)}
\]

\[
1 - \sum_{i=1}^{N} a_i z^{-i} \quad \text{if} \quad i=1 \quad N
\]

(4.3.22)

Any intermediate value of \( S \) between 0 and 1 achieves a different degree of noise shaping. For optimum perceptual effect, \( S \) is typically around the value of 0.6.

The three noise shaping configurations discussed in this and the previous sections produce an output noise with a pole-zero spectrum. The set of poles of the noise spectrum is the same as that of the speech spectrum. They all achieve equivalent noise shaping function with slightly different configurations. The salient feature in the pre- and post-filtering arrangement is the need to transmit the coefficients of the pre-filter in addition to that of the spectral envelope predictor \( A_s(z) \). The pre-filter also removes part of the redundancy in the speech signals by performing partial prediction. The combined prediction gain of \( R(z) \) and \( A(z) \) also means a smaller value of \( S_q(f) \) in Equation (4.3.21) compared to that in Equation (4.3.5) in the previous configuration.
Figure 4.3.8  Block diagram of a generalised predictive coder using
pre- and post-filtering to achieve the desired spectrum
of the output noise

U.Q. = Unit variance quantizer
Finally, it would be incomplete without mentioning the use of fixed pre- and de-emphasis filter to produce output noise with a pole spectrum. The value of the coefficient of the pre- and de-emphasis filter is adjusted to give the best possible perceptual effect. Its effectiveness in producing better perceptual quality speech will be compared with that of the adaptive noise shaping schemes.

4.3.5 Backward Adaptive Noise Spectral Shaping

The technique of adaptive feedback filtering to achieve noise shaping described in the previous sections requires the buffering of typically 16 msec of speech samples for the calculation of the coefficients for both the predictor and the noise feedback filter. The coefficients are then quantized and transmitted to the receiver as side-information. In the case of noise shaping via pre-filtering an additional set of coefficients for the pre-filter also has to be transmitted. The delay and transmission of side information are obviously the two major drawbacks of these systems. The straightforward approach to overcome these problems is to replace all the forward adaptive elements in the system by backward adaptive elements. Thus, the forward adaptive quantizer (AQF) has to be replaced by a backward adaptive quantizer like AQJ. The forward adaptive predictor has to be replaced by a backward adaptive predictor which derives its prediction coefficients from the previous block of locally decoded samples. The coefficients for the noise feedback filter and the pre-filter can in turn be derived from the prediction coefficients.
Though a backward ANS technique can avoid the problems of delay and transmission of side-information, it is not as effective as the forward adaptation schemes which have the ability to 'look ahead'. Another disadvantage of the backward ANS scheme is the possibility of 'mistracking' when transmission errors occur. When the transmitted bit stream is corrupted by noise in the channel, the locally decoded samples at the receiver are different from those at the transmitter. Consequently, the coefficients of the predictor and the noise feedback filter or pre-filter at the receiver are different from those at the transmitter. Unless some mechanism is incorporated into the system to dissipate the effect of transmission errors, the filters can diverge and operate independently of one another.

The use of backward ANS technique for the coding of narrowband speech at 16 kbps was investigated by F. Yeoh and C. Xydeas (137). The perceptual quality of the coded speech was found to be enhanced by the use of the backward ANS technique. Subjectively, it was equivalent to 7-bit PCM.
4.4 ADPCM CODING OF WIDEBAND SPEECH WITH AND WITHOUT NOISE SHAPING AT 32 kbps

The basic system block diagram of an ADPCM coder is shown in Figure 4.4.1. The predictor used in the system is a 4th-order Forward Block Adaptive Predictor (FBAP). For every 256 input sample (≈ 16 msec) of speech, the prediction coefficients are calculated using the auto-correlation method and updated accordingly. The quantizer is the Adaptive Quantizer Forward (AQF) with uniform step-size. The step-size $\Delta$ is also updated every 16 msec and obtained by first calculating the RMS value $G$ of the residue given by

$$G = \text{RMS} \left[ S(z)(1 - A(z)) \right]$$

(4.4.1)

and then multiplying $G$ by a constant of 1.8, i.e.

$$\Delta = 1.8 G$$

(4.4.2)

The constant 1.8 was obtained through computer optimization using speech samples as the input. Obviously, the 4 prediction coefficients and $G$ have to be multiplexed together with the speech samples as side information and transmitted to the receiver. An additional 1.9 kbps is therefore incurred as 6 bits are required every 16 msec to quantize each coefficient and the step-size. The total bit rate of the system is therefore equal to 33.9 kbps. The system is abbreviated as ADPCM/AQF/FBAP.

To examine the effectiveness of noise spectral shaping applied to 2-bit coarse quantization ADPCM coder, the following four ADPCM with noise spectral shaping configurations NS1, NS2, NS3 and NS4 were simulated.
(NS1): ADPCM with fixed first order pre-emphasis and de-emphasis as shown in Figure 4.4.2. It is abbreviated as ADPCM/AQF/FBAP/PDE.

(NS2): ADPCM with Makhoul's noise spectral shaping configuration. The 4th-order adaptive noise shaping filter is incorporated into the feedback loop of the system. A block diagram of the system is given in Figure 4.4.3. It is abbreviated as ADPCM/AQF/FBAP/ANS.

(NS3): This system employs the adaptive noise shaping technique of (NS2) with the addition of pre-emphasis and de-emphasis filter. The system block diagram is given in Figure 4.4.4. It is abbreviated as ADPCM/AQF/FBAP/PDE/ANS.

(NS4): This system employs 4th-order adaptive pre- and post-filtering. The system block diagram is shown in Figure 4.4.5. It is abbreviated as ADPCM/AQF/FBAP/PPF.

The bit rates of NS1, NS2 and NS3 are the same as that of ADPCM/AQF/FBAP of 33.9 kbps. As for NS4, an additional 1.5 kbps of side information is required for the transmission of the four pre-filter coefficients. This adds up to a total of 35.4 kbps for the system.
Calculate $A(Z)$

Figure 4.4.1 The ADPCM/AQF/FBAP configuration

Calculate Predictor Coefficients $(a_1, a_2, a_3, a_4)$

Figure 4.4.2 The ADPCM/AQF/FBAP/PDE configuration (NSl)

$C(Z) = 1 - 0.8 Z^{-1}$
Figure 4.4.3 The ADPCM/AQF/FBAP/ANS configuration (NS2)

Figure 4.4.4 The ADPCM/AQF/FBAP/ANS/PDE configuration (NS3)

Figure 4.4.5 The ADPCM/AQF/FBAP/PPF configuration (NS4)
4.5 RESULTS AND DISCUSSIONS

Computer simulations of the various noise spectral shaping schemes were performed using the four speech sentences listed in Section 3.6 as input data. Evaluation and comparisons of the performances of the systems were carried out using average SNRSEG measurements, long-term power spectral density plots of the output noise and informal subjective listening tests.

4.5.1 Results and Comparisons of the ADPCM and NS1 Systems

The coefficient of the pre- and de-emphasis filter in the NS1 system was set to 0.8 which was found to give the best perceptual quality for the recovered speech. The average segmental SNR measurements of the ADPCM and NS1 systems are given in Table 4.5.1. The overall average SNRSEG of NS1 is lower than that of the ADPCM by 1.2 dB. This suggests that the noise accumulation effect of the de-emphasis filter outweighs the advantage gained by the dynamic range reduction of the input signal due to pre-emphasis.

The long-term average output noise spectra of the two systems are given in Figure 4.5.1. For the ADPCM system, the average output noise above the frequency of 3.7 kHz is above the speech signals by a considerable extent. This suggests the presence of high frequency 'hissing' noise in the recovered speech, which was confirmed by subjective listening tests.
<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>ADPCM/AQF/FBAP (in dB)</th>
<th>ADPCM/AQF/FBAP/PDE (NS1) (in dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.72</td>
<td>29.97</td>
</tr>
<tr>
<td>2</td>
<td>27.87</td>
<td>26.52</td>
</tr>
<tr>
<td>3</td>
<td>22.91</td>
<td>21.47</td>
</tr>
<tr>
<td>4</td>
<td>19.26</td>
<td>16.96</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>24.94</strong></td>
<td><strong>23.73</strong></td>
</tr>
</tbody>
</table>

Table 4.5.1 SNRSEG measurements for the ADPCM/AQF/FBAP and NS1 systems.

Figure 4.5.1 Long-term average power spectral density plots of:
(1) Original speech
(2) Output noise of ADPCM/AQF/FBAP
(3) Output noise of ADPCM/AQF/FBAP/PDE (NS1)
For the NS1 system, the energy of the output noise at the higher part of the spectrum is considerably reduced but at the expense of increasing the energy of the output noise at the lower part of the spectrum. The increased low frequency 'rumbling' noise is still effectively masked by the speech formants for three of the sentences. The masking is however unsatisfactory for the second female sentence. Nevertheless, with great reduction in high frequency 'hissing' noise, the NS1 system is perceptually better than the ADPCM system although its SNRSEG value is lower than the latter.

4.5.2 Results of the NS2 System

In the NS2 system, the adaptive noise spectral shaping configuration was implemented in the way suggested by Makhoul. No pre- and de-emphasis was used. It was therefore a test of the effectiveness of the adaptive noise spectral shaping technique applied to coarse quantization (2 bit/sample) ADPCM coder without the help of pre- and de-emphasis. The noise shaping filter \( 1 - B(z) \) has the following transfer function

\[
1 - B(z) = 1 - \frac{A(z/B)}{A(z)} \tag{4.5.1}
\]

where \( 0 < B < 1 \).

A male sentence of 4.2 seconds was used to test the effect of adaptive noise shaping for different values of \( B \). When the value of \( B \) was varied from 0.4 to 0.9, the SNRSEG value increased from 14.33 dB to
22.53 dB (see Table 4.5.2 and Figure 4.5.2). Figure 4.5.3 shows the output noise spectra for different values of $\beta$. As the value of $\beta$ decreases, the shape of the output noise spectrum follows closer to that of the original speech spectrum. This is apparent in the low frequency region. The output noise in the higher part of the spectrum is also slightly reduced. Notice that the shaping is not very pronounced as only a 4th-order shaping filter is used. A higher order shaping filter ($\geq 8$) will theoretically achieve finer spectral shaping but that will mean an increase in system complexity and overall transmission bit rate for the system because more coefficients have to be transmitted to the receiver.

Perceptually, a value of $\beta = 0.5$ or below was found to be unacceptable for the sentences used in the simulations as the low frequency output noise was too high to be effectively masked by the speech signal. A $\beta$ value of 0.6 seemed to be the best choice for the test sentences. Any value higher than that implies mild noise shaping and the perceptual improvement was found to be negligible.
Table 4.5.2: SNRSEG measurements for the ADPCM/AQF/FBAP/ANS system (NS2) with different values of $\beta$ for sentence No. 3.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRSEG (dB)</td>
<td>14.33</td>
<td>15.80</td>
<td>17.70</td>
<td>19.54</td>
<td>21.24</td>
<td>22.53</td>
<td>22.91</td>
</tr>
</tbody>
</table>

Table 4.5.3: SNRSEG measurements for the ADPCM/AQF/FBAP/ANS system (NS2) with $\beta = 0.6$ for 4 different sentences

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.02</td>
</tr>
<tr>
<td>2</td>
<td>20.60</td>
</tr>
<tr>
<td>3</td>
<td>17.70</td>
</tr>
<tr>
<td>4</td>
<td>15.70</td>
</tr>
<tr>
<td>Average</td>
<td>19.00</td>
</tr>
</tbody>
</table>
Figure 4.5.2  SNRSEG performance of the ADPCM/AQF/FBAP/ANS system (NS2) against different values of $\beta$

Figure 4.5.3  Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the ADPCM/AQF/FBAP/ANS (NS2) system for $\beta = 0.5$
(3) $\beta = 0.6$
(4) $\beta = 0.7$
4.5.3 Comparisons of the ADPCM, NS1 and NS2 Systems  
ADPCM vs. NS2

The difference between ADPCM and NS2 is that the latter has an additional adaptive noise shaping filter incorporated into the feedback loop. As discussed in the previous section, the value of $\beta = 0.6$ was found to be the best noise shaping constant for the NS2 system. Therefore, the comparison will be between the ADPCM and the NS2 system with $\beta = 0.6$. (See Figure 4.5.4, Tables 4.5.1 and 4.5.3).

In terms of SNRSEG, ADPCM has a higher overall average value of 24.94 dB compared to 19.00 dB of NS2. Subjectively, the recovered speech of NS2 is slightly more pleasing than that of ADPCM. Though the high frequency 'hissing' noise of NS2 is reduced compared to that in ADPCM, it is still slightly above the speech signal and perceptually objectionable. Therefore, the overall subjective improvement due to the use of ANS technique is not very significant at this bit rate. This agrees with Bastian's\(^{135}\) observation that for noise shaping to be effective, the system must have reasonably good quality to begin with. Without pitch prediction or entropy encoding\(^{131,132}\), coarse quantization ADPCM does not seem to be able to exploit fully the technique of adaptive noise shaping.

It is to be noted that our results do not necessarily contradict the observation done by Makhoul\(^{132}\). In his paper he reported that ANS
produced excellent quality when used in ADPCM coding of narrowband speech at 16 kbps. This is due to the fact that entropy encoding and a 19-level quantizer were employed. Consequently, the quantization noise was lower than the ADPCM scheme examined and thus the job of the ANS filter was made very much easier and its effect was more pronounced.

**NS1 vs. NS2**

The SNRSEG of NS1 has an average value of 23.73 dB compared to 19.00 dB for NS2. The subjective quality of the two systems are quite different. The main subjective degradation of NS1 is in the low frequency region. The non-adaptive de-emphasis filter of NS1 increase the magnitude of low frequency noise and the masking effect provided by the speech signal is not completely satisfactory at this bit rate. However, the high frequency 'hissing' noise is very successfully depressed. For the NS2 system, the main source of degradation comes from the high frequency region. Though high frequency quantization noise is reduced to a certain degree in NS2, it is still above the speech signal and perceptually unacceptable. However, the low frequency degradation in NS2 is lower than that in NS1 due to the adaptive nature of ANS. A plot of the output noise spectra of both systems in Figure 4.5.5 gives an objective description of the nature of the noise produced by them.
Figure 4.5.4  Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of ADPCM/AQF/FBAP
(3) Output noise of ADPCM/AQF/ANS (NS2) for $\beta = 0.6$

Figure 4.5.5  Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of ADPCM/AQF/FBAP/PDE (NS1)
(3) Output noise of ADPCM/AQF/FBAP/ANS (NS2) for $\beta = 0.6$
4.5.4 Results of the NS3 System

In the NS1 and NS2 systems, noise spectral shaping is accomplished through the use of the PDE and ANS filters respectively. The NS3 system (ADPCM/AQF/FBAP/ANS/PDE) combines the two noise shaping techniques of NS1 and NS2. The speech signal is first pre-emphasised by a fixed first order pre-filter of transfer function $C(z) = 1 - 0.8 z^{-1}$ before being encoded by the ADPCM/ANS system.

The value of $\beta$ that determines the transfer function of the adaptive noise shaping filter was re-optimized with respect to the pre-emphasised speech. Subjectively, the best choice of $\beta$ was found to be 0.9 compared to 0.6 for the use of speech without pre-emphasis. This is hardly surprising as a greater degree of noise shaping (i.e. $\beta < 0.9$) coupled with the de-emphasis will give rise to a very high level of low frequency noise and renders the quality of the recovered speech objectionable. This also means that there is not much room for noise shaping when PDE is used.

Table 4.5.4 shows the SNRSEG measurements of the system for different values of $\beta$. As $\beta$ increases from 0.7 to 0.9, SNRSEG increases from 15.39 to 19.36 dB. When $\beta = 1.0$, the system is equivalent to NS1 and the SNRSEG value is 21.47 dB. The power spectral density plots of the output noise for different values of $\beta$ are given in Figure 4.5.6. The SNRSEG measurements for the system with $\beta = 0.9$ for the four different sentences are also given in Table 4.5.5.
Table 4.5.4  SNRSEG measurements for the ADPCM/AQF/FBAP/ANS/PDE system (NS3) for different values of $\beta$ for sentence No. 3.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRSEG(dB)</td>
<td>15.39</td>
<td>17.27</td>
<td>19.36</td>
<td>21.47</td>
</tr>
</tbody>
</table>

Table 4.5.5  SNRSEG measurements for the ADPCM/AQF/FBAP/ANS/PDE system (NS3) with $\beta = 0.9$ for the 4 different sentences.

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>26.23</td>
</tr>
<tr>
<td>2</td>
<td>23.37</td>
</tr>
<tr>
<td>3</td>
<td>19.36</td>
</tr>
<tr>
<td>4</td>
<td>15.90</td>
</tr>
<tr>
<td>Average</td>
<td>21.22</td>
</tr>
</tbody>
</table>

Figure 4.5.6  Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of ADPCM/AQF/FBAP/ANS/PDE for $\beta = 0.8$
(3) $\beta = 0.9$
(4) $\beta = 1.0$
Subjectively, the NS3 system is only marginally better than the NS1 system. Its perceptual quality is similar to that of NS1 because both systems employ PDE and consequently have very little high frequency 'hissing' noise. In comparison with NS2, it has a clear advantage because the former has unacceptable degradation in the high frequency region.
4.5.5 Results of the NS4 System

The last noise shaping configuration employs an adaptive pre-filter with $\beta = 0.6$ to produce a partially spectrally flattened signal $y_n$ which is then encoded by the ADPCM/AQF/FBAP system. The choice of $\beta$ was optimized to obtain the best perceptual quality for the output speech. The inverse filtering process performed by the adaptive post-filter at the receiver shapes the quantization noise so that it has a time-varying pole-zero spectrum determined by the set of poles of the speech spectrum. It is therefore functionally equivalent to the NS2 system in the aspect of noise shaping. The difference between them is that the shaping of noise in NS4 is carried out on the noise which has a lower power level than that in NS2 due to the partial prediction gain obtained by pre-filtering.

The NS4 system is also a generalized system of NS1. When $\beta = 0$ and the order of the pre-filter is reduced to 1 with the coefficient set to 0.8, the adaptive pre-filter in NS4 becomes the pre-emphasis filter in NS1.

The SNRSEG measurements for the NS4 system for different values of $\beta$ are given in Table 4.5.6. Once again, SNRSEG increases with the increase in $\beta$. When $\beta = 1$, the ADPCM, NS2 and NS4 systems are equivalent and give the same SNRSEG measurement. For other values of $\beta$, SNRSEG measurements of NS4 are higher than that of NS2 as shown in Figure 4.5.6. Subjectively, it also has a lower level of perceived
output noise than that in NS2. The perceptual improvement is significant but at the cost of an increase in bit rate due to the need to transmit the adaptive pre-filter coefficients.

In comparison with NS1, the subjective improvement of NS4 was noticeable especially in the 0-3 kHz region. However, NS1 has a slightly lower level of noise than NS4 in the higher part of the spectrum. Spectral comparison of the output noise of the NS1, NS2 and NS4 systems are shown in Figure 4.5.7.
Table 4.5.6  SNRSEG measurements for the ADPCM/AQF/FBAP/PPF system (NS4) with different value of $\beta$ for sentence No. 3.

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRSEG(dB)</td>
<td>16.72</td>
<td>18.31</td>
<td>19.96</td>
<td>21.40</td>
<td>22.91</td>
</tr>
</tbody>
</table>

Table 4.5.7  SNRSEG measurements for the ADPCM/AQF/FBAP/PPF system (NS4) with $\beta = 0.6$ for the 4 different sentences

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG(dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.89</td>
</tr>
<tr>
<td>2</td>
<td>23.03</td>
</tr>
<tr>
<td>3</td>
<td>18.31</td>
</tr>
<tr>
<td>4</td>
<td>16.83</td>
</tr>
<tr>
<td>Average</td>
<td>20.52</td>
</tr>
</tbody>
</table>
Figure 4.5.6 SNRSEG performances of the NS2 and NS4 systems against different values of $\beta$.

Figure 4.5.7 Long-term average power spectral density plots of:
1. Output noise of the NS2 system ($B = 0.6$)
2. Output noise of the NS4 system ($B = 0.6$)
3. Output noise of the NS1 system
4.6 NOTE ON PUBLICATION

A paper entitled "Noise Spectral Shaping Applied to Coarse Quantization Differential Speech Coders" was presented at the Mediterranean Electrotechnical Conference (MELECON '83) in May 1983 and was recorded in the Conference Proceedings pp. C1.08. It was written in co-authorship with Dr. C. S. Xydeas and Dr. F. Yeoh and covers the work described in this chapter.

4.7 SUMMARY AND CONCLUSION

The traditional design of high bit rate speech coder like the 64 kbps PCM coding of narrowband speech is based primarily on minimizing the quantization noise power in the recovered speech. The spectral shape of the quantization noise is relatively flat with respect to that of the speech spectrum. For medium and low bit rate speech coder design, SNR is no longer the only design criterion. The spectral shape of the output noise spectrum has to be adjusted in relation to that of the speech spectrum so that the effect of the auditory masking properties of the ear can be exploited in order to produce a perceptually more pleasing quality of recovered speech. This Chapter has presented a study of the performance of fixed and adaptive noise spectral shaping techniques applied to coarse quantization ADPCM speech coder at 2 bits/Nyquist sample. The simple noise spectral shaping takes the form of a fixed first order pre- and de-emphasis. The adaptive noise spectral shaping schemes studied include the backward adaptive feedback of
quantization noise and the forward adaptive pre- and post-filtering technique.

Computer simulations revealed that the forward adaptive pre- and post-filtering technique offers the best possible perceptual improvement for ADPCM coding of wideband speech at a bit rate slightly higher than 32 kbps. However, the recovered speech is still slightly degraded by a low but marginally audible level of quantization noise. The improvement obtained therefore does not really justify the high complexity involved in calculating the pre-filter and prediction coefficients. Furthermore, the coefficients have to be transmitted as side-information to the receiver.

The other noise spectral techniques examined were also found to enhance the perceptual quality of the ADPCM coded speech. The use of the simple fixed first order pre- and de-emphasis is very effective in reducing the high frequency 'hissing' noise present in the ADPCM coded speech at 2 bits/sample, though at the expense of an increase in low frequency noise. Auditory masking of the low frequency noise by the speech spectrum is partially satisfactory. The perceptual improvement due to such a simple noise shaping procedure suggests that it is advantageous to include such a procedure in the design of an ADPCM coder.

The technique of adaptive filtering of the quantization noise via a feedback loop in an ADPCM coder, proposed by Atal and Makhoul, to achieve
adaptive noise spectral shaping was also found to be effective, but not to a satisfactory degree, in improving the perceptual quality of the ADPCM coded speech. This shows that the quantization noise level has to be sufficiently low for any noise shaping technique to be fully effective. Under coarse quantization, the simple technique of pre- and de-emphasis is perhaps the most cost-effective means to improve the subjective quality of the recovered speech. All the above observations suggest that adaptive noise spectral shaping of the quantization noise is an important issue not to be neglected in the design of a time-domain waveform coder at medium bit rate, it does not give much improvement.

In the next two chapters, the more effective frequency-domain waveform coders like subband coder and adaptive transform coder will be examined in detail for the coding of wideband speech at 32 kbps. Unlike time-domain waveform coders, SBC and ATC schemes have the inherent advantage that not all the samples need to be quantized with equal number of bits. The allocation of bits for the quantization of the frequency components or transform coefficients can be based on perceptual criterion.
CHAPTER 5

SUBBAND CODING (SBC) TECHNIQUES
5.1 INTRODUCTION

The use of the various adaptive noise spectral shaping techniques has been shown in the previous chapter to improve the subjective quality of the recovered speech in the ADPCM coding of wideband speech at 2 bits per sample. However, the overall perceptual quality of the decoded speech was found to be below that produced by the 2-band or the ADPCM/lattice prediction coder operating at 64 kbps. This is due to the fact that the quantization noise in a 2-bit ADPCM coder is just not low enough for the coder to fully exploit the promise of the ANS techniques.

The other possible solution to achieve good quality wideband speech at 2 bits per sample is either the subband coding (SBC) or the adaptive transform coding (ATC) technique in the frequency-domain waveform coding category. By transforming the speech signals in the time-domain into components in the frequency-domain via either the filter-bank analysis or the block transformation techniques, quantization of the components based on perceptual criterion can be achieved by allocating more bits to the perceptually more significant components and less bits or no bit for those components which do not contribute much to the overall subjective quality. Furthermore, a higher number (8 and above) of subbands or larger transform blocksize implies greater frequency resolution and hence finer control of output noise in relation to the speech spectrum. On top of that, subband coding has the added advantage of the confinement of quantization
noise within each band and thus preventing the masking of the speech signal in one frequency range by the quantization noise in another frequency range. The price paid is the higher system complexity and extra delay incurred by employing these techniques.

In this chapter, the various aspects pertaining to the design of a subband coder are first described. The theories of the conventional Quadrature Mirror Filter (QMF)\(^{(73)}\) and the less known Complex Quadrature Mirror Filter (CQMF)\(^{(152)}\) are presented. The derivation of a novel bit allocation algorithm is proposed as an alternative to the conventional adaptive bit allocation algorithm, in order to simplify the implementation of a subband coder.

The performances of three groups of subband coders are examined in this chapter. The first group consists of the seven-band coders employing the various forward and backward quantization strategies and three different bit allocation algorithms. The use of pre- and post-filtering to enhance the quality of the recovered speech produced by a seven-band SBC coder is also considered. The second group of SBC coders are similar to that in the first group except that the number of bands is increased to fourteen. The performance of the coders under the various transmission error conditions is examined also. The last group of coders employs CQMF to achieve division of the full band speech into seven uniform pairs of amplitude and phase signals. Instead of using the adaptive bit allocation
or the simplified bit allocation algorithm, bit allocation patterns are vector quantized into seven patterns and employed in the design of a complex subband coder (CSBC). A novel quantization strategy which exploits the correlations between the amplitude signals from adjacent subbands is proposed. For the quantization of the phase signals, a simplified bit allocation algorithm which derives the allocation patterns from the decoded amplitude signals, is employed.

The performances of all the three groups of subband coders are assessed and compared in terms of average segmental SNR and average spectral density plots of the output noise. Informal subjective listening tests were carried out also as a final performance criterion.
5.2 SPEECH CODING IN SUBBANDS

In the category of waveform coding of speech, quantization noise is the main source of degradation of the quality of the decoded speech. Quantization is a non-linear process and produces quantization noise that is typically broad in frequency spectrum. Therefore, for speech segments with mainly low frequency content, high frequency quantization noise is audible and for speech segments with high frequency content, low frequency quantization noise can also be perceptually annoying.

Quantization noise is not equally detectable at all frequencies because of the characteristics of the speech spectrum and the masking effect it provides. There are several techniques of reducing the effect of quantization noise to improve the quality of the recovered speech. The noise shaping techniques described in the previous chapter and the noise rejection technique (153-154) proposed by Jayant and Smith are two possible techniques one can apply. Another way of achieving this objective is to encode speech in its subbands rather than the coding of the full band signals.

Basically, subband coding (SBC) is the splitting of speech into a number of different frequency subbands followed by waveform encoding of the signals in each subband. The straightforward approach is to carry out frequency translation of the high frequency subbands into basebands, down sample each subband output at its Nyquist rate and
encode the signals in each subband with the same or different types of encoding techniques and finally, multiplex all the encoded subband signals together with some side-information, if there is any, for transmission. At the receiver, the subband signals are decoded and bandpass translated back to their original spectral position. They are then summed to give a replica of the original speech signal.

5.2.1 Division of Speech Spectrum into Subbands

The number of subbands and the width of each band are two related design issues in the division of speech spectrum in a subband coding system. Typically, the full band signal is divided into 4 or more subbands (68-70). For low bit rate transmission, if the number of subbands is less than 4, there is little advantage to be gained over the full band coder. On the other hand, if the number of subbands is more than 4, system complexity increases. With the advent of more powerful digital signal processing devices, there is a tendency to use more than 4 subbands in the design of a SBC coder. In fact the use of a 32-band SBC coder for the coding of narrowband speech at 16 kbps and below had been reported by Crochiere (79).

The partitioning of the full band speech signal can be based on the concept of Articulation Index (AI). The Articulation Index AI was defined by Beranek (13) as "a number obtained from articulation tests using nonsense syllables under the assumption that any narrowband of speech frequencies of a given intensity in the absence of noise
carries a contribution to the total index, which is independent of the other bands with which it is associated, and that the total of all the bands is the sum of the contributions of the separate bands". The articulation test is used as a quantitative measure of the intelligibility of speech transmitted over a communication system. Thus, the full band speech signal is divided in such a way that each band contributes equally to the AI. Alternatively, the full band signal can be partitioned into uniform bandwidth subbands and the technique of integer-band sampling can be exploited to achieve lowpass translation (LPT) simply by decimation \((68,155)\).

When the AI concept is the chosen design criterion, the bandwidth of each subband would be non-uniform with the lower bands occupying narrower bandwidths than the upper bands. However, each band is equally important perceptually. One example of dividing narrowband speech into 4 subbands using the concept of AI is given below:

<table>
<thead>
<tr>
<th>Subband No.</th>
<th>Frequency Range (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200 - 700</td>
</tr>
<tr>
<td>2</td>
<td>700 - 1310</td>
</tr>
<tr>
<td>3</td>
<td>1310 - 2020</td>
</tr>
<tr>
<td>4</td>
<td>2020 - 3200</td>
</tr>
</tbody>
</table>

The drawback of the AI design criterion is that the technique of integer-band sampling cannot be used to achieve LPT. A more complicated modulation procedure is required to translate each of the higher frequency subbands to the baseband. Another disadvantage of designing according to the AI criterion is the high order \((> 200)\) of finite impulse response.
(FIR) transversal filter required to achieve sharp cut-off in order to avoid aliasing due to the down sampling of the signal in each band. Aliasing in the frequency domain manifests itself in the form of reverberation effect in the reconstructed speech and proves to be perceptually unacceptable.

If the division into subbands is uniform across the full band, the technique of Quadrature Mirror Filtering can be employed. Band partitioning in this case is accomplished by using a symmetrical filter tree structure, with each branch of the tree consisting of FIR filters of theoretically any even order, without giving rise to the problem of frequency aliasing in the reconstructed signal. A more detailed discussion on QMF is given in Section 5.2.3.

5.2.2 Low Pass Translation (LPT) of Subband Signals

The very first step in processing subband signals before encoding is to lowpass translate the signals to the baseband so that the sampling rate can be reduced to the Nyquist rate of that band. One of the methods to achieve LPT is straightforward modulation by a cosine wave with frequency equal to the lower cut-off sequency $W_{in}$ of the particular band. The modulated signal is filtered by a lowpass filter $h_n(t)$ with bandwidth $(0 - W_n)$ equal to that of the particular subband to remove the unwanted signal images above $2W_{in}$. The whole process is best illustrated in Figure 5.2.1. The lowpass translated signal $r_n(t)$ for the nth band can be expressed as
where $s_n(t)$ is the bandpass signal of the nth subband. It can be easily seen that there is a restriction of $\omega_n \leq 2W_n$ imposed by this method. The signal $r_n(t)$ is down-sampled to its Nyquist rate $2W_n$, digitally encoded and multiplexed with the encoded signals from other subbands and transmitted to the receiver. At the receiver, the data is de-multiplexed into separate subbands, decoded and interpolated to give the estimate signal $\tilde{r}_n(t)$ for the nth band. To obtain the estimate $s_n(t)$ of the original nth band signal $s_n(t)$, $\tilde{r}_n(t)$ is modulated by $\cos (\omega_n t)$ and bandpass filtered to its original spectral range.

To avoid the above mentioned restriction on $\omega_n$, the technique of complex modulation can be employed. The nth band signal $s_n(t)$ is modulated by both $\cos \omega_n t$ and $\sin \omega_n t$ in two branches where $\omega_n$ is the centre frequency of the nth band. The modulated signals in the two branches are then lowpass filtered to the bandwidth of $(0 - W_n/2)$ to yield $a_n(t)$ and $b_n(t)$. The signals $a_n(t)$ and $b_n(t)$ are further modulated by $\cos (W_n t/2)$ and $\sin (W_n t/2)$ respectively and their sum is the lowpass translated signal $r_n(t)$ of $s_n(t)$. Detailed discussion on complex modulation can be found in reference (68). Whether it be straightforward or complex modulation, there is a need of real number multiplication of the subband signals by sinusoidal signals in order to achieve LPT.
Figure 5.2.1  Sequence of operations of low pass translation of a speech subband and band restoration
To avoid the need of real number modulation in the process of lowpass translation, the very simple integer-band sampling technique can be employed instead. The full band speech signal is divided into $N$ equally spaced subbands with $mf$ as the lower cut-off frequency and $(m+1)f$ as the upper cut-off frequency for the $(m+1)$th band. Each subband signal is then down sampled to $2f$ and the resulting frequency aliased signal is actually equivalent to the lowpass translated signal of that band. At the receiver, each subband signal is recovered by decoding and bandpassing the decoded signal to its original spectral position. The spectral interpretation of this technique is shown in Figure 5.2.2.
Figure 5.2.2  Integer-band sampling technique for digital encoding of speech subbands
5.2.3 Quadrature Mirror Filter (QMF)

In the earlier designs of subband coding systems, large order (200 tap or more) sharp cut-off FIR filters were required to minimize the effects of frequency aliasing due to the down sampling of the subband signals. To eliminate this requirement, D. Estaban proposed the use of the elegant design of Quadrature Mirror Filter (73) to achieve band division. The order of FIR filter used in the QMF analysis bank can then be reduced to 32 or less without the problem of aliasing.

Consider the two-band division system shown in Figure 5.2.3. Let $H_1(\omega)$ be the Fourier transform of the impulse response of the lowpass filter with cut-off frequency equal to $1/4$ of the sampling frequency $\omega_s$. If the high pass filter $H_2(\omega)$ with cut-off frequency of $1/4 \omega_s$ is the half-band mirror image of $H_1(\omega)$, then

$$H_1(e^{j\omega}) = H_2(e^{j(\omega_s/2 - \omega)}) \quad (5.2.2)$$

Similarly, assume that $K_1(\omega)$ and $K_2(\omega)$ be the LPF and HPF of the same cut-off frequency at the receiver respectively and $K_2(\omega)$ is the mirror image of $K_1(\omega)$.

The output signals of the filters $H_1$ and $H_2$ are down sampled by 2 : 1 decimation to give $y_1(z)$ and $y_2(z)$. The Z transform of the decimated signals can be expressed as

$$y_1(z) = \frac{1}{2} \{X(Z^{1/2}) H_1(Z^{1/2}) + X(-Z^{1/2})H_1(-Z^{1/2})\} \quad (5.2.3)$$

$$y_2(z) = \frac{1}{2} \{X(Z^{1/2}) H_2(Z^{1/2}) + X(-Z^{1/2})H_2(-Z^{1/2})\}$$
Two-band division and reconstruction using Quadrature Minor Filter (QMF)

Spectral interpretation of $H_1$ and $H_2$
For reconstruction, $y_1$ and $y_2$ are 2:1 up-sampled by inserting zero between every two samples and filtered by $K_1$ and $K_2$ before being added together to give the reconstructed signal $S$, i.e.

$$S(Z) = T_1(Z) + T_2(Z)$$

(5.2.4)

where

$$T_1(Z) = \frac{1}{2} \{ X(Z)H_1(Z) + X(-Z)H_1(-Z) \} K_1(Z)$$

(5.2.5)

$$T_2(Z) = \frac{1}{2} \{ X(Z)H_2(Z) + X(-Z)H_2(-Z) \} K_2(Z)$$

Combining Equations (5.2.4) and (5.2.5) gives

$$S(Z) = \frac{1}{2} \{ H_1(Z)X_1(Z) + H_2(Z)X_2(Z) \} X(Z) +$$

$$\frac{1}{2} \{ H_1(-Z)X_1(Z) + H_2(-Z)X_2(Z) \} X(-Z)$$

(5.2.6)

The second term of Equation (5.2.6) represents the aliasing effect due to down sampling. It can be eliminated with proper choice of $K_1$, $K_2$, $H_1$ and $H_2$. One restriction on the choice of the filters is that $H_2$ is the half band mirror image of $H_1$ and $K_2$ is the mirror image of $K_1$, i.e. Equation (5.2.2) must be satisfied. This condition can be easily met if $H_1(Z)$ and $H_2(Z)$ are FIR filters given by

$$H_1(Z) = \sum_{n=0}^{N-1} h_1(n)Z^{-n}$$

(5.2.7)

and

$$H_2(Z) = \sum_{n=0}^{N-1} h_1(n)(-1)^n Z^{-n} = H_1(-Z)$$

(5.2.8)
The filter $H_2$ is obtained by inverting the sign of every other sample of the impulse response of $H_1$. This process of sign inversion of $H_1$ can be interpreted as modulating a signal with the frequency spectrum of $H_1$ by a sinusoidal wave of frequency $\frac{1}{2} \omega_s$ and thus forming a half band mirror image of $H_1$. When $H_2(Z)$ is defined by Equation (5.2.8), the second term of Equation (5.2.6) can be eliminated if

$$K_1(Z) = H_1(Z)$$  \hspace{1cm} (5.2.9)

and

$$K_2(Z) = -H_2(Z) = -H_1(-Z)$$  \hspace{1cm} (5.1.10)

The reconstructed signal is then given by

$$S(Z) = \frac{1}{2} \{H_1^2(Z) - H_1^2(-Z)\} X(Z)$$  \hspace{1cm} (5.2.11)

When expressed in the frequency domain, $S(Z)$ becomes

$$S(e^{j\omega T}) = \frac{1}{2} \{H_1^2(e^{j\omega T}) - H_1^2(e^{j(\omega + \frac{\omega_s}{2}) T})\} X(e^{j\omega T})$$  \hspace{1cm} (5.2.12)

If $H_1$ is a symmetrical FIR filter, i.e.

$$h_1(n) = h_1(N-1-n), \ n = 0, 1, 2, \ldots, N/2 - 1$$  \hspace{1cm} (5.2.13)

The Fourier transform $H_1(e^{j\omega T})$ can be expressed in terms of its amplitude response $H_1(\omega)$ as

$$H_1(e^{j\omega T}) = H_1(\omega) e^{-j(N-1)\pi \frac{\omega}{\omega_s}}$$  \hspace{1cm} (5.2.14)
where the phase response of \( e^{-j \frac{(N-1)2\pi}{\omega_s}} \) arises from the pure time delay of \( \frac{(N-1)}{2} \) samples introduced by the symmetrical transversal filter. Equation (5.2.12) now becomes

\[
S(e^{j\omega T}) = \frac{1}{2} \left( H_1^2(\omega) - H_1^2(\omega + \frac{\omega_s}{2}) \right) e^{-j(N-1)\pi} \cdot e^{-j(N-1)2\pi \frac{\omega}{\omega_s}} \cdot X(e^{j\omega T})
\]  
\[ (5.2.15) \]

When \( N \) is odd, it can be easily seen from Equation (5.2.15) that the reconstructed signal \( S \) is always equal to zero at \( \omega = \frac{\omega_s}{4} \). Therefore \( N \) can only be even in order to recover the signal. For \( N \) equals to even, \( S(e^{j\omega T}) \) reduces to

\[
S(e^{j\omega T}) = \frac{1}{2} \left( H_1^2(\omega) + H_1^2(\omega + \frac{\omega_s}{2}) \right) e^{-j(N-1)2\pi \frac{\omega}{\omega_s}} \cdot X(e^{j\omega T})
\]  
\[ (5.2.16) \]

The reconstructed signal is thus the original signal \( x \) with an amplitude distortion of \( \left( H_1^2(\omega) + H_1^2(\omega + \frac{\omega_s}{2}) \right) \), which can be made negligible and perceptually unnoticeable, and a pure delay of \( N-1 \) samples. The amplitude distortion can be minimized by the proper design of \( H_1 \). By using Hooks and Jeaves optimization routine (156) with a Hanning window prototype, Johnson had successfully designed a filter family for the use in QMF banks (157). Figure 5.2.5 illustrates the mechanism of aliasing cancellation in a two-band analysis and synthesis process.
Figure 5.2.5 Spectral Interpretation of Two-band Division and Reconstruction using QMF (See Figure 5.2.3).
To recapitulate, the reconstructed signal is aliasing free if the following conditions are met:

1) $H_1(Z)$ is symmetrical FIR filter of even order

2) $H_2(Z) = H_1(-Z)$

3) $K_1(Z) = H_1(Z)$

4) $K_2(Z) = -H_1(-Z)$

To divide the full band into four or more subbands, the same process is repeated at the lower and upper subband and the resulting filter bank becomes a tree structure. Alternatively, Cheung had shown that it is possible to implement the QMF bank in a parallel structure (76). QMF bank can also be implemented using recursive IIR filters (158,159). However, its performance is inferior to that obtained using FIR filters due to the non-linear phase distortion of IIR filter.

In speech analysis and synthesis using a symmetrical QMF tree structure, perfect cancellation of aliasing due to the down-sampling of the subband signals can be achieved. However, it also implies rigid uniform division of the full speech band, and thus the concept of A.I. cannot be applied.

If the QMFs are used to form a partial tree structure as shown in Figure (5.2.6), non-uniform and octavely spaced subband division more in line with the division on A.I. can be achieved. In this case, cancellation of aliasing is no more perfect. Pre-emphasis of the input
Figure 5.2.6 Use of QMFs in a partial tree structure to achieve the division of full band speech into octavely spaced subbands.
speech signals before the analysis filter bank\(^{(77)}\) is sometimes used to flatten the speech spectrum and consequently cuts down the amount of energy spilling or aliasing inherent in this kind of asymmetrical structure. The aliasing components are hoped to be masked by the speech signals. This design strategy may not be suitable for high quality speech coders because to begin with, aliasing distortion is unavoidable no matter how good the quantization process is.
5.2.4 **Complex Quadrature Mirror Filter (CQMF)**

The technique for the decomposition of a signal into N adjacent subbands uniformly spaced in the frequency domain using QMFs can be extended to decompose a signal into N adjacent pairs of complex channels using complex QMF (CQMF)\(^{(160,161)}\). Consider a signal \(x(t)\) to be one of the N subbands of a QMF filter bank. Let the bandwidth of \(x(t)\) be \(\omega_s/2\)\(^{\left(= 2\pi f_s/2\right)}\) and its sampling frequency be \(f_s\)\(^{\left(= \frac{1}{T}\right)}\). To convert \(x(t)\) into two quadrature components, first modulate \(x(t)\) in two branches by two sinusoidal signals of frequency \(f_s/4\) and a phase difference of \(\pi/2\) (see Figure 5.2.7). Since the sampling frequency is \(f_s\), the modulating carrier signals are simply

\[
\cos(2\pi f_s nT/4) = 1, 0, -1, 0, \ldots \quad \text{and} \\
\sin(2\pi f_s nT/4) = 0, 1, 0, -1, \ldots
\]

The modulation procedure can therefore be easily implemented.

Next, the two modulated signals are passed through two M-tap FIR low-pass filters each with a bandwidth of \(f_s/4\) and transfer function \(H(Z)\) to give two signals \(\hat{y}(n)\) and \(y(n)\). In Z-transform notation,

\[
y(Z) = \frac{1}{2} \left[ X(jZ) + X(-jZ) \right] H(Z) \\
\hat{y}(Z) = \frac{1}{2j} \left[ X(jZ) - X(-jZ) \right] H(Z)
\]

(5.2.17)

The two signals \(y(n)\) and \(\hat{y}(n)\) can now be down-sampled by 2:1 decimation to give two quadrature signals \(v(n)\) and \(\hat{v}(n)\).
Figure 5.2.7 Decomposition of a signal into (a) $N$ pairs of complex channels and (b) a pair of complex channels using CQMF.
To reconstruct the original signal $x(n)$, the two quadrature signals $v(n)$ and $\hat{v}(n)$ are upsampled by inserting zero between every two samples and lowpass filtered by $H(Z)$ followed by demodulation with $\sin(2\pi f_s nT/4)$ and $\cos(2\pi f_s nT/4)$ respectively. Finally, the reconstructed signal $\hat{x}(n)$ is obtained by subtracting the two demodulated signals. The final expression for $\hat{x}(n)$ in $Z$ transform notation can be shown (Appendix III) to be given by

$$\hat{x}(Z) = \frac{1}{4j} X(Z) \left[ H^2(jZ) - H^2(-jZ) \right]$$

(5.2.19)

It shows that the final reconstructed signal $\hat{x}(n)$ is free from aliasing which corresponds to the term $X(-Z)$.

Assume that $H(Z)$ is a symmetrical FIR filter of order $M = 4K$ where $K$ is a positive integer and $H(\omega)$ is the magnitude of the Fourier transform $H(e^{j\omega T})$. The Fourier transform of $x(n)$ can be shown from Equation (5.2.19) to be given by (Appendix III).

$$\hat{x}(e^{j\omega T}) = \frac{1}{4} X(e^{j\omega T}) e^{-j(M-1)\omega T} \left[ H^2(\omega + \frac{\omega_s}{4}) + H^2(\omega - \frac{\omega_s}{4}) \right]$$

(5.2.20)

With the proper design of $H$,

$$H^2(\omega + \frac{\omega_s}{4}) + H^2(\omega - \frac{\omega_s}{4}) = 1$$
and \( \hat{X}(e^{j\omega T}) \) becomes

\[
\hat{X}(e^{j\omega T}) = \frac{1}{4} X(e^{j\omega T}) e^{-j(M-1)\omega T}
\]  

(5.2.21)

The reconstruction of \( x(n) \) is therefore almost perfect except a delay of \( M-1 \) samples and a scaling factor of \( \frac{1}{4} \).

When this decomposition and reconstruction procedure is applied to the \( N \) subband signals of a QMF filter bank, we obtain \( N \) pairs of complex signals with the guarantee of aliasing free synthesis at the receiver. The \( N \) pairs of complex signals can be used to form \( N \) pairs of amplitude \( A \) and phase \( \phi \) signals according to the equations

\[
A(n) = \sqrt{\nu^2(n) + \nu^2(\hat{\nu}(n))}
\]

\[
\phi(n) = \tan^{-1} \frac{\nu(\hat{\nu}(n))}{\nu(n)}
\]

(5.2.22)

It thus provides another alternative form of signal representation which may lead to a more efficient coding technique.
5.2.5 Fixed and Adaptive Bit Allocation Algorithms (FBA/ABA)

A very important feature in subband coding is the non-uniform allocation of bits among the subbands to control the amount of quantization noise in each band to achieve a perceptually improved performance. The non-flat characteristics of speech spectrum and the different degree of masking of the quantization noise at different parts of the frequency spectrum suggest that the allocation of bits should make full use of these properties. When fixed bit allocation (FBA) is chosen for a SBC system, the long-term perceptual criteria of each subband should be the basis of allocation. In this case, more bits are reserved for the lower frequency bands where pitch and formant structure must be more accurately preserved than for the upper frequency bands where speech components are more noise like. Typically, the number of bits allocated to the lower bands is about 4 and to the upper bands is either 1 or 2. The possibility of allocating different numbers of bits to the subbands implicitly implies that the spectral shape of the noise can be controlled to a certain degree across the full band.

As the short-term frequency spectra of speech vary greatly from one segment to another, fixed bit allocation is clearly not an optimum strategy. When a coding system can afford higher complexity, adaptive bit allocation (ABA) can be employed to take into account the time-varying short-term speech spectrum. The design of the bit allocation algorithm will determine whether the system has the best SNR measurement or the best perceptual performance.
In general, for subband coding employing an ABA algorithm, the transmission of speech is divided into frames of 16 to 20 msec. The variance $\sigma_i^2$ of the signal in each band within each frame is calculated and used to derive the bit allocation pattern for that frame. If an average mean-squared distortion $D_i$ is not to be exceeded for the quantization of the $i$th band signal with variance $\sigma_i^2$, the number of bits $R_i$ assigned for the quantization of each sample in that band is given by the well known equation (71)

$$R_i = \delta + \frac{1}{2} \log_2 \frac{\sigma_i^2}{D_i}, \quad i = 1, \ldots, N. \quad (5.2.23)$$

$N$ is the total number of bands and $\delta$ is the correction value that takes into account the performance of practical quantizers. The second term of the equation is closely related to the PCM equation (2.3.10). From the PCM equation, the allocation of one additional bit leads to a 6 dB improvement in the quantization process. 6 dB is equivalent to a signal-to-noise ratio, $\sigma_i^2/D_i$, of 4 which means the second term of equation (5.2.23) is equal to 1. In fact, the PCM equation can be obtained by re-arranging equation (5.2.23).

Let the average distortion be $\overline{D}$ and the constant average bit rate be $\overline{R}$, i.e.

$$\overline{D} = \frac{1}{N} \sum_{i=1}^{N} D_i \quad (5.2.24)$$

and

$$\overline{R} = \frac{1}{N} \sum_{i=1}^{N} R_i = \text{constant} \quad (5.2.25)$$
The optimum bit assignment for all the subband signals to have the same distortion \( D_i = \bar{D} \) for all \( i \) can be easily shown to be

\[
R_i = \bar{R} + \frac{1}{2} \log_2 \left( \frac{\sigma_i^2}{\prod_{j=1}^{N} \sigma_j^2} \right) \quad \text{(bits/sample)} \tag{5.2.26}
\]

These results were first derived by Huang and Schulthesis\(^{(162)}\). In the application of adaptive bit allocation using equation (5.2.26), the value of \( R_i \) is usually rounded to the nearest integer or set to zero when \( R_i \) turns out to be negative. The maximum rounded value of \( R_i \) is also restricted to a certain value in practical implementation. These operations may result in the total number of bits allocated less than or more than the number of bits available for transmission. In this case, additional number of bits can be added or deleted arbitrarily from the subbands\(^{(71)}\). Alternatively, the bands that receive zero or negative bit allocation are ignored and the same algorithm repeated for the remaining bands\(^{(71)}\).

Though the adaptive bit allocation algorithm of Equation (5.2.26) maximizes the SNR performance of the coder, it is not necessarily the most desirable because the output noise of the coder has a relatively flat spectrum. If perceptual quality is the design criterion, the output noise of the system has to be spectrally shaped with respect to the speech spectrum. To achieve noise spectral shaping, the bit assignment algorithm is modified by scaling each \( \sigma_i^2 \) by a weighting
factor \( W_i \). With this modification, Equation (5.2.26) becomes

\[
R_i = \bar{R} + \frac{1}{2} \log_2 \left( \frac{W_i \sigma_i^2}{\left[ \sum_{j=1}^{N} W_j \sigma_j^2 \right]^{1/N}} \right) \tag{5.2.27}
\]

The weighting factor \( W_i \) that was found to give the desired noise shaping properties is given by

\[
W_i = \sigma_i^{2\gamma} \tag{5.2.28}
\]

The parameter \( \gamma \) can be experimentally adjusted between 0 and -1 to give the best perceptual result.

When \( \gamma = -1 \), \( R_i \) is equal to \( \bar{R} \) for all \( i \) and the noise spectrum follows closely the speech spectrum. Thus, 'complete' noise spectral shaping is achieved. This corresponds to \( \beta = 0 \) in the case of ADPCM with adaptive noise shaping configuration (Equation 4.3.19).

When \( \gamma = 0 \), Equation (5.2.27) reduces to Equation (5.2.26) and the output noise spectrum is flat. This corresponds to \( \beta = 1 \) in the ADPCM/ANS system. Intermediate values of \( \gamma \) between 0 and -1 will produce different degree of noise spectral shaping. By using Equation (5.2.28), the bit assignment Equation (5.2.27) can be re-written as

\[
R_i = \bar{R} + \frac{1}{2} \log_2 \left( \frac{\sigma_i^{2(1+\gamma)}}{\left[ \sum_{j=1}^{N} \sigma_j^{2(1+\gamma)} \right]^{1/N}} \right) \tag{5.2.29}
\]
There are other simplified bit allocation algorithms (163) which produce results close to that obtained by using Equation (5.2.29). A novel simplified bit allocation algorithm will also be described in the following section.

5.2.6 Simplified Bit Allocation Algorithm (SBA)

The problem with the bit allocation strategy of Equation (5.2.29) is that the value of $R_i$ is usually fractional and sometimes negative. The process of rounding $R_i$ to an integer value and setting negative $R_i$ to zero may result in the total number of bits allocated exceeding or falling below the number of bits available. Hence re-adjustments have to be made to ensure that the coder is operating at the correct bit rate. To avoid these problems and the processing requirements of Equation (5.2.29), the following simplified adaptive bit allocation algorithm is proposed.

Equation (5.2.29) is used to determine the bit allocation patterns for all the 16 msec frames in an input training sequence which is free of any silent interval. The value of $\gamma$ is set to -0.3 and the maximum number of bits $M$ allowed in the allocation is set to 5. Let $N_i$ represent the total number of times that $i$ bits are allocated in the training sequence and $i = 1, \ldots, 5$. Next, we define $f_i$ as

$$f_i = \frac{N_i \cdot i}{N_T} \quad i = 1, \ldots, 5$$

(5.2.30)
where \( N_T = \sum_{i=1}^{M} N_i \cdot i \).

\( N_T \) is the total number of bits available for allocation throughout the training sequence. \( f_i \) therefore represents the portion of \( N_T \) used in allocating \( i \) bits for the coding of subband signals. If there are \( N_T \) bits available for allocation in each frame, the expected number of bands \( n_i \) that receive \( i \) bits can be calculated according to

\[
n_i = \text{nearest integer of } \left( \frac{f_i \cdot N_T}{i} \right) \tag{5.2.31}
\]

for \( i = 1, \ldots, M \).

A time-invariant bit allocation pattern (IBAP) is thus obtained using the \( n_i \) estimates, i.e.:

\[
\{ n_5 \times 5 \text{ bits}, n_4 \times 4 \text{ bits}, \ldots, n_1 \times 1 \text{ bit}, 0 \text{ bit for the remaining bands} \}
\]

This means that, within a 16 msec frame, \( n_5 \) subbands receive 5 bits, \( n_4 \) subbands receive 4 bits and so on. Manual adjustment is normally required to ensure that the total number of bits in the IBAP pattern gives the correct transmission bit rate.

Though the pattern is fixed, the allocation is based on the step-sizes of the subband signals. For each frame of 16 msec, the \( n_5 \) subbands with the first \( n_5 \) largest step-sizes are allocated 5 bits each and the next \( n_4 \) subbands with the next \( n_4 \) largest step-sizes are allocated 4 bits each and so on.
The use of this simplified bit allocation algorithm (SBA) requires the transmission of the quantizer step-sizes of the subband signals to the receiver. This is also required when using the fully adaptive bit allocation algorithm of Equation (5.2.29). Compared to ABA, the processing requirement of SBA is almost trivial. The use of SBA is not restricted to subband coding. Any system employing the ABA strategy can use the SBA as a substitute to reduce system complexity. Of course there would be a drop in system performance due to the use of the SBA algorithm instead of the ideal ABA algorithm. However, the drop in system performance might be perceptually insignificant and the SBA algorithm could prove to be a better alternative in terms of system complexity. The various IBAP patterns for the 7 and 14-band SBC coders and an ATC coder will be given in the subsequent sections. Their performance will be compared with the coders employing the conventional ABA algorithm.
5.2.7 Vector Quantization of Bit Allocation Patterns

The simplest form of a subband coding scheme employs fixed bit allocation for the quantization of the subband signals. The bit allocation pattern is designed to suit the long-term average spectral characteristics of the speech signals. When the designer can afford higher implementation complexity, the fully adaptive bit allocation algorithm is usually used for the system to adapt to the time-varying short-term spectra of speech. The use of the full ABA algorithm leads to an infinite number of bit allocation patterns which may not be essential as far as the perceptual quality of the recovered speech is concerned. In between these two extremes, the bit allocation patterns may be vector quantized into a finite number of fixed allocation patterns.

Vector quantization is essentially a clustering technique (34, 35, 164, 165) which groups a collection of multi-dimensional vectors into a finite number of clusters each of which can be used to obtain a single vector to represent vectors that fall within that cluster. The grouping of vectors is based on certain criterion like the minimum Euclidean distance measure. As the bit allocation patterns are derived from the variance vectors calculated every short segment of speech, the vector quantization of bit patterns is equivalent to the vector quantization of the variance vectors. The normalized variance vector $\Lambda$ is defined as

$$\Lambda = \left(\frac{\sigma_1^2}{\sigma_T^2}, \frac{\sigma_2^2}{\sigma_T^2}, \frac{\sigma_3^2}{\sigma_T^2}, \ldots, \frac{\sigma_N^2}{\sigma_T^2}\right)$$

(5.2.32)
where \[ \sigma_T^2 = \sum_{i=1}^{N} \sigma_i^2 \]

and \( N \) is the total number of subbands and \( \sigma_i^2 \) represents the variance of the subband signals of the \( i \)th-band within a short frame of speech. Normalization is necessary because speech power varies from segment to segment. A collection of \( \Lambda \)'s is obtained from a training sequence which does not contain silent intervals between speech bursts. The conventional technique of vector quantization \(^{(164)}\) is then applied to this collection of vectors to obtain a finite number of variance vector \( V_i \) using the minimum Euclidean distance criterion. From the vectors \( V_i \)'s, the finite number of fixed bit allocation patterns can be derived using Equation (5.2.29).

For the application of the vector quantized bit allocation patterns in a coding system like a SBC or ATC coder, the finite number (usually small) of \( V_i \)'s are stored both at the receiver and transmitter. The bit allocation pattern is then determined by measuring the distance between the transmitted variance vector \( \Lambda \) for a frame of speech and the stored vectors \( V_i \)'s. The bit allocation pattern that corresponds to the vector \( V_i \) that gives the minimum distance will be employed for that frame of speech. When the vector quantization of the bit allocation pattern is correctly done, one would expect some patterns to correspond to the voiced speech and the others to correspond to unvoiced speech. The use of this particular form of bit allocation strategy is further discussed in Sections (5.5).
5.3 **SEVEN-BAND SUBBAND CODERS (7/SBC)**

Since the conception of the idea of encoding independently the subband instead of the full band signals of speech, various subband coding schemes with different number of subbands and using different adaptive bit allocation algorithms and quantization strategies have received considerable attention for the coding of narrowband speech at 16 kbps or below for various applications (73-83). However, only its simplest form of 2-band coding with fixed or adaptive bit allocation has been investigated for the coding of wideband speech at 64 kbps (80,81,120-123, 146,147,166). To achieve good quality recovered speech at 32 kbps, the number of bands has to be increased in order to exploit more effectively the advantages of quantization noise confinement and the flexibility of different bit assignments for the subband signals.

The first group of subband coders for the coding of wideband speech at 32 kbps are the seven-band subband coders (7/SBC) with the following different configurations:

1. 7/SBC with fixed bit allocation and uniform AQFs.
2. 7/SBC with fixed bit allocation and ADPCM/AQJ encoders.
3. 7/SBC with adaptive bit allocation and uniform AQFs.
4. 7/SBC with simplified bit allocation and Laplacian AQFs.
5. 7/SBC with adaptive pre- and post-filtering, fixed bit allocation and AQFs.

* As a four-band subband coder is not expected to produce the required commentary-grade quality of recovered speech, our investigation begins with seven-band (eight bands minus the last band) subband coders.
5.3.1 Seven-Band Subband Coding with Fixed Bit Allocation and Uniform AQFs

The basic system block diagram of a tree-structured QMF seven-band subband coder is shown in Figure 5.3.1. QMF's are used to divide the full band speech signal into seven uniform bands. The prototype lowpass filter is the 32-tap FIR filter designed by Johnston(81). Table 5.3.1 contains the impulse response of the filter. Its frequency response is shown in Figure 5.3.2. The total delay, in terms of number of samples \( N_d \), of the input signal introduced by the filter banks can be easily shown to be given by

\[
N_d = (N - 1) + 2(N - 1) + ... + 2^{P-1} (N - 1)
\]

(5.3.1)

\[
= (N - 1)(2^P - 1)
\]

where \( N \) is the order of the prototype FIR used and \( P \) denotes the number of stages in the filter bank. As the 32-tap FIR filter is used throughout all the stages in the transmitter and receiver and \( P \) is equal to 3 for a 7-band SBC coder, the total delay is 217 samples. With a sampling frequency of 16 kHz, it is equivalent to 13.56 msec. When signal delay is a vital system design constraint, lower-order FIR filters can be used for the second and higher stages of the tree-structured filter bank to reduce the total delay introduced by them.

For the quantization of the subband signals, Adaptive Quantizers Forward (AQFs) are employed. AQF is also known alternatively as block companded PCM(BCPCM)(73). It is basically equivalent to linear PCM.
Figure 5.3.1 The system block diagram of a tree-structured QMF seven-band subband coder
<table>
<thead>
<tr>
<th>h(1)</th>
<th>0.224500E-02 = h(32)</th>
</tr>
</thead>
<tbody>
<tr>
<td>h(2)</td>
<td>-0.397100E-02 = h(31)</td>
</tr>
<tr>
<td>h(3)</td>
<td>-0.197000E-02 = h(30)</td>
</tr>
<tr>
<td>h(4)</td>
<td>0.818200E-02 = h(29)</td>
</tr>
<tr>
<td>h(5)</td>
<td>0.843000E-03 = h(28)</td>
</tr>
<tr>
<td>h(6)</td>
<td>0.142290E-01 = h(27)</td>
</tr>
<tr>
<td>h(7)</td>
<td>0.206900E-02 = h(26)</td>
</tr>
<tr>
<td>h(8)</td>
<td>0.227040E-01 = h(25)</td>
</tr>
<tr>
<td>h(9)</td>
<td>-0.796200E-02 = h(24)</td>
</tr>
<tr>
<td>h(10)</td>
<td>-0.349640E-01 = h(23)</td>
</tr>
<tr>
<td>h(11)</td>
<td>0.194720E-01 = h(22)</td>
</tr>
<tr>
<td>h(12)</td>
<td>0.548120E-01 = h(21)</td>
</tr>
<tr>
<td>h(13)</td>
<td>-0.445240E-01 = h(20)</td>
</tr>
<tr>
<td>h(14)</td>
<td>-0.993390E-01 = h(19)</td>
</tr>
<tr>
<td>h(15)</td>
<td>0.132973E+00 = h(18)</td>
</tr>
<tr>
<td>h(16)</td>
<td>0.463674E+00 = h(17)</td>
</tr>
</tbody>
</table>

Table 5.3.1 Impulse response of the 32-tap FIR transversal filter $H_1$

Figure 5.3.2 Frequency response of the 32-tap FIR filter $H_1$
or uniform quantizer with its step-size updated every short time interval of typically 16 to 20 msec because speech signals are generally stationary within this interval. The optimum step-size $\Delta$, calculated for every block of 16 msec of subband signals, is given by the equation

$$\Delta = \alpha \left( \frac{1}{N} \sum_{i=1}^{N} x_i^2 \right)^{1/2}$$

(5.3.2)

where $N = 32$ for the subband signal sampled at 2 kHz and $x_i$'s are the sample values of a particular band. $\alpha$ is the scaling constant obtained through computer optimization for the best SNR performance. It varies with the number of quantization levels and the pdf of the input signals. Table 5.3.2 contains the values of $\alpha$ for different number of quantization levels for each subband signal. The step-sizes of the seven AQFs have to be quantized and multiplexed with the encoded subband signals for transmission. Five bits are generally sufficient for the quantization of each quantizer step-size and therefore an additional 2.2 kbps is required for the 7/SBC coder. The total system bit rate is now 34.2 kbps. As AQF introduces another 16 msec of delay, the total system delay is 29.56 msec.

For the fixed bit allocation scheme, four fixed allocation patterns given in Table 5.3.3 were experimented on. No bit is allocated to encode the last subband signal for the 4th assignment pattern. However,
the rms value $\sigma_7$ of the samples of that band is calculated and transmitted every 16 msec to the receiver. A Gaussian signal with zero mean and a standard deviation of $0.8\sigma$ is then generated and used as a substitute input for that band at the receiver. Examples of the time waveforms of the seven subband signals are given in Figure 5.3.3a to 5.3.3g.
<table>
<thead>
<tr>
<th>Subband (kHz)</th>
<th>Number of bits</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>(1) 0 - 1</td>
<td>0.94</td>
</tr>
<tr>
<td>(2) 1 - 2</td>
<td>1.00</td>
</tr>
<tr>
<td>(3) 2 - 3</td>
<td>1.04</td>
</tr>
<tr>
<td>(4) 3 - 4</td>
<td>1.06</td>
</tr>
<tr>
<td>(5) 4 - 5</td>
<td>1.02</td>
</tr>
<tr>
<td>(6) 5 - 6</td>
<td>1.04</td>
</tr>
<tr>
<td>(7) 6 - 7</td>
<td>1.04</td>
</tr>
</tbody>
</table>

Table 5.3.2 The value of $a$ for different subbands and different quantizer levels

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>Band No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4 3 2 2 2 2 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5 4 3 1 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5 4 2 2 1 1 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6 4 2 2 1 1 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3.3 The four fixed bit allocation patterns
Figure 5.3.3a - 5.3.3g The time waveforms of the output signals of the seven subbands. (Sampling frequency = 2 kHz)
5.3.1.1 Results and Discussions

The operation of the seven-band subband coder with the four fixed bit allocation patterns were simulated using as input the four different speech data files described in Chapter 3. Coder performance was evaluated in terms of average segmental SNR measurements, long-term average spectral density plots of the output noise and informal subjective listening tests. Figures 5.3.4 and 5.3.5 show the long-term average power spectral plots for the 7/SBC coder employing the four different bit allocation patterns. As revealed by the spectral plots and confirmed by informal listening tests, the first pattern of (4, 3, 2, 2, 2, 2, 1) and the second pattern of (5, 4, 3, 1, 1, 1, 1) did not give acceptable performance as there were insufficient numbers of bits allocated to the first two bands in the former and too few bits were allocated to the 4th band in the latter. The third assignment pattern of (5, 4, 2, 2, 1, 1, 1) gave perceptually more pleasing results than the first two patterns. However, a slight 'coarseness', due to the marginally high level of quantization noise in the low frequency region, is still perceptible. In the 4th bit allocation pattern, the lowest band receives 6 bits at the expense of the highest band. The pattern is now (6, 4, 2, 2, 1, 1, 0). The long-term average spectral plot of the output noise of the coder employing the 4th pattern and informal subjective listening tests showed that there was an improvement in the first band but more distortion, in the form of high frequency noise, was perceived in the last band. Thus, there was no clear cut preference for either of the last two bit allocation patterns.
Figure 5.3.4  Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of 7/SBC/AQF coder employing the 1st FBA pattern
(3) Output noise of 7/SBC/AQF coder employing the 2nd FBA pattern
(4) Output noise of 7/SBC/AQF coder employing the 3rd FBA pattern

Figure 5.3.5  Long-term average spectral density plots of
(1) Original speech
(2) Output noise of 7/SBC/AQF coder employing the 3rd FBA pattern
(3) Output noise of 7/SBC/AQF coder employing the 4th FBA pattern
The average segmental SNR measurements of the coders employing the 3rd and 4th bit allocation patterns are given in Table 5.3.4. The 7/SBC coder employing the 3rd bit allocation pattern has an average SNR performance of 21.50 dB compared to 24.44 dB of the same coder employing the 4th bit allocation pattern. However, this big difference of 3 dB in SNR performance corresponds only to a marginal difference in the perceptual quality of the recovered speech.

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SBC(3rd FBA) dB</th>
<th>SBC(4th FBA) dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>24.61</td>
<td>29.01</td>
</tr>
<tr>
<td>2</td>
<td>22.99</td>
<td>26.68</td>
</tr>
<tr>
<td>3</td>
<td>20.36</td>
<td>22.68</td>
</tr>
<tr>
<td>4</td>
<td>18.02</td>
<td>19.39</td>
</tr>
<tr>
<td>Average</td>
<td>21.50</td>
<td>24.44</td>
</tr>
</tbody>
</table>

Table 5.3.4 SNRSEG measurements for the subband coding scheme with fixed bit allocation
5.3.2 Seven-band Subband Coding with Fixed Bit Allocation and ADPCM/AQJ Encoders

The seven-band subband coder employing fixed bit allocation and AQFs requires an additional 2.2 kbps for the transmission of step-sizes as side-information. One way of reducing this amount of side-information is to employ differential encoding with adaptive Jayant's quantizer. It will also be of interest to examine how well differential encoding with fixed prediction and Jayant's quantizer performs for subband signals in comparison with AQF. The system configuration is described in Table 5.3.5.

The first five bands employ ADPCM coding with fixed first or second order prediction. The bit allocation is fixed to (5, 3, 2, 2, 2, 1, 1). AQF's are still retained for the last two bands. The average rms value $\sigma_{67}$ of the last two bands' signal is calculated and transmitted as side-information to the receiver. At the receiver, the step-size value of the 5-6 kHz band is set to 1.2 $\sigma_{67}$ and that of the 6-7 kHz band is set to 0.8 $\sigma_{67}$. The scaling constants are set in accordance with the observation that the long-term speech spectrum has a slightly higher power level for the 6th band than the 7th band. This avoids the transmission of two step-size values. The bit rate is thus reduced to 32.3 kbps.

A second experiment of seven-band subband coding employing ADPCM/AQJ encoders was also performed, similar to the scheme just described but
with pre-emphasis and de-emphasis of 0.4 for the first band. This is an attempt to shape the first band noise spectrum in order to obtain perceptually improved performance. The prediction coefficients for the pre-emphasised first band signal are $a_1 = 0.4121$ and $a_2 = -0.5202$.

<table>
<thead>
<tr>
<th>No.</th>
<th>Subband (kHz)</th>
<th>Bit Allocation</th>
<th>Encoder</th>
<th>Predictor Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0 - 1</td>
<td>5</td>
<td>ADPCM/FSOP</td>
<td>$a_1 = 0.6931, a_2 = -0.5811$</td>
</tr>
<tr>
<td>2</td>
<td>1 - 2</td>
<td>3</td>
<td>ADPCM/FFOP</td>
<td>$a_1 = -0.0864$</td>
</tr>
<tr>
<td>3</td>
<td>2 - 3</td>
<td>2</td>
<td>&quot;</td>
<td>$a_1 = 0.0600$</td>
</tr>
<tr>
<td>4</td>
<td>3 - 4</td>
<td>2</td>
<td>&quot;</td>
<td>$a_1 = -0.1458$</td>
</tr>
<tr>
<td>5</td>
<td>4 - 5</td>
<td>2</td>
<td>&quot;</td>
<td>$a_1 = 0.1625$</td>
</tr>
<tr>
<td>6</td>
<td>5 - 6</td>
<td>1</td>
<td>AQF</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>6 - 7</td>
<td>1</td>
<td>&quot;</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.3.5 System configuration for the subband coder with FBA and ADPCM/AQJ
5.3.2.1 Results and Discussions

The long-term average spectral density plots of the 7/SBC schemes employing ADPCM/AQJ encoders are shown in Figure 5.3.6 where they are compared with that of the 7/SBC/AQF scheme which employs the 3rd bit allocation pattern. It can be seen that the quantization noise level is increased in all the bands except for the 5th band. Subjective listening tests confirmed this objective indication that the 7/SBC coder which employs AQFs outperforms the same coder with ADPCM/AQJ. This is hardly surprising as there is little correlation between adjacent samples in each subband for the differential coders to exploit. The second 7/SBC/ADPCM scheme with pre- and de-emphasis for the first band gave perceptually slightly lower level of low frequency distortion due to the fact that quantization noise in the first band was shaped to follow the speech spectrum and thus more effective masking of noise by the speech signals was achieved.

The average SNRSEG values (see Table 5.3.6) for the two systems are 20.98 dB and 20.51 dB with the one employing PDE having the lower value. This compares not too badly with the 21.50 dB and 24.44 dB of the 7/SBC/AQF coder employing the 3rd and 4th bit allocation patterns respectively, despite the fact that side-information is cut down from 2.2 kbps to 0.3 kbps.
Table 5.3.6 SNRSEG measurements for the two subband coding with differential encoders and FBA pattern schemes

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SBC/ADPCM/FBA</th>
<th>SBC/ADPCM/PDE/FBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.27</td>
<td>22.51</td>
</tr>
<tr>
<td>2</td>
<td>24.69</td>
<td>24.29</td>
</tr>
<tr>
<td>3</td>
<td>18.34</td>
<td>17.86</td>
</tr>
<tr>
<td>4</td>
<td>17.60</td>
<td>17.38</td>
</tr>
<tr>
<td>Average</td>
<td>20.98</td>
<td>20.51</td>
</tr>
</tbody>
</table>

5.3.6 Long-term average power spectral density plots of

(1) Original speech
(2) Output noise of 7/SBC/ADPCM/AQJ
(3) Output noise of 7/SBC/ADPCM/AQJ/PDE
(4) Output noise of 7/SBC/AQF coder employing the 3rd FBA pattern
5.3.3 Seven-Band Subband Coding with Adaptive Bit Allocation and Uniform AQFs

The SNR and subjective performance of a subband coder can be enhanced by employing an adaptive bit allocation strategy to efficiently assign the correct number of bits to the required subbands as the frequency characteristics of speech vary from frame to frame. Typically, speech is considered stationary within a time-frame of 16 msec. The bit allocation pattern of a coder is therefore updated every 16 msec to track the varying short-term speech spectrum.

The bit allocation algorithm given by Equation (5.2.29) was employed in the simulation of the operation of a 7/SBC/ABA coder. The step-sizes of the AQFs were calculated and transmitted every 16 msec to the receiver for the quantization of the subband signals. As the subband signals' variances required for the determination of the bit allocation patterns, were the squares of the step-sizes, no additional side-information was incurred by employing ABA strategy. The number of bits allocated to a certain band can exceed six when applying the ABA equation for certain segment of speech. As it was found perceptually unnecessary to have more than six bits allocated to any one band, the maximum number of bits allowed was clamped to six. When the algorithms gave negative bit allocation to a certain band for some speech segments, no bit was assigned to that band. Instead, the transmitted step-size $\Delta$ for that band was used to generate a Gaussian signal with zero mean and a standard deviation of $0.8 \Delta$ which was
inserted into that band to avoid the creation of spectral gaps in the reconstructed signal.

As the ABA algorithm often produced non-integer values, a rounding operation was required to arrive at integer bit allocation. This rounding operation together with the setting of negative bit allocation to zero and clamping of maximum number of bits allowed to six often resulted in the total number of bits allocated in a frame exceeding or falling below the number of bits available for quantization. When more bits than those available were allocated, the extra bits were taken away, each from the subbands with the least step-sizes. If less bits than available were allocated the remaining bits were added to the subbands with the largest step-sizes provided the maximum allocation for these bands did not exceed six.

5.3.3.1 Results and Discussions

The effectiveness of the ABA algorithm given by Equation (5.2.29) in the 7/SBC coding of wideband speech was first tested with the value of γ set to zero. Figure 5.3.7 shows a reduction in the level of noise in the 2-3 kHz and 4-5 kHz bands for the 7/SBC/ABA scheme compared to the 7/SBC/FBA scheme with the third and forth patterns. It also achieved a theoretically predicted flat noise spectrum except for the first band which had a slightly higher level of noise. This was because the maximum number of bits allowed was clamped to six.
As predicted by the noise spectral plot, the subjective performance of the 7/SBC/ABA scheme was better than that of the 7/SBC/FBA schemes. Average segmental SNR measurements (see Table 5.3.7) showed also a higher average value of 26.59 dB compared to 21.51 dB and 24.44 dB for the 7/SBC/AQF coder employing the third and fourth FBA patterns. The price paid, of course, was the additional complexity of implementing the ABA algorithm.

The use of the ABA algorithm with $\gamma = 0$ produces relatively flat noise spectrum which is not the most desirable perceptually, though SNR performance is maximized. By varying the value of noise shaping parameter $\gamma$, the output noise can be controlled to a certain degree as shown in Figure 5.3.8. Redistribution of the power of the output noise from the higher part of the spectrum to the lower part of the spectrum can be observed with decreasing values of $\gamma$. The segmental SNR performance of the coder decreases as $\gamma$ decreases from 0 to -0.5 (see Table 5.3.8). Subjectively, $\gamma = -0.5$ produced too much low frequency noise. The best perceptual performance seemed to be produced by the coder with $\gamma = -0.3$. With this value of $\gamma$, the average SNRSEG was measured to be 25.43 dB (see Table 5.3.9) compared to 26.59 dB when $\gamma = 0.0$. 
Table 5.3.7  SNRSEG measurements for the 7/SBC/ABA ($\gamma = 0$) scheme

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29.90</td>
</tr>
<tr>
<td>2</td>
<td>28.13</td>
</tr>
<tr>
<td>3</td>
<td>25.58</td>
</tr>
<tr>
<td>4</td>
<td>22.74</td>
</tr>
<tr>
<td>Average</td>
<td>26.59</td>
</tr>
</tbody>
</table>

Figure 5.3.7  Long-term average power spectral density plots of
(1) Original speech
(2) Original noise of 7/SBC/AQF coder employing the 3rd FBA pattern
(3) Output noise of 7/SBC/AQF coder employing the 4th FBA pattern
(4) Output noise of 7/SBC/ABA ($\gamma = 0$) coder
<table>
<thead>
<tr>
<th>$\gamma$</th>
<th>-0.5</th>
<th>-0.4</th>
<th>-0.3</th>
<th>-0.2</th>
<th>0.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNRSEG (dB)</td>
<td>21.64</td>
<td>23.64</td>
<td>24.40</td>
<td>25.04</td>
<td>25.58</td>
</tr>
</tbody>
</table>

Table 5.3.8: SNRSEG measurements of the 7/SBC/ABA scheme with different values of $\gamma$

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>28.61</td>
</tr>
<tr>
<td>2</td>
<td>27.09</td>
</tr>
<tr>
<td>3</td>
<td>24.40</td>
</tr>
<tr>
<td>4</td>
<td>21.63</td>
</tr>
<tr>
<td>Average</td>
<td>25.43</td>
</tr>
</tbody>
</table>

Table 5.3.9 SNRSEG measurements of the 7/SBC/ABA Scheme for $\gamma = -0.3$

Figure 5.3.8 Long-term average power spectral density plots of

(1) Original Speech
(2) Output noise of the 7/SBC/ABA coder with $\gamma = -0.3$
(3) $\gamma = -0.2$
(4) $\gamma = -0.1$
(5) $\gamma = 0.0$
5.3.4 Seven-Band Subband Coding with Simplified Bit Allocation and Laplacian AQFs

The use of adaptive bit allocation in the coding of wideband speech at 2 bits/sample seems to be essential as it produces both higher SNR performance and improved quality for the recovered speech compared to the same coder employing fixed bit allocation. Furthermore, it does not incur additional bit rate as the bit assignment patterns are derived from the transmitted step-sizes of the AQFs of the subbands. The only problem with ABA is the additional processing requirement of Equation (5.2.29). \( \log_2 \) look up table is required in the determination of the bit allocation patterns. Furthermore, rounding operation described in the previous section leads to an additional re-adjustment procedure in order to ensure that the total number of bits allocated is equal to the number of bits available. To avoid all these problems and yet maintain the advantage of adaptive bit allocation, the simplified bit allocation algorithm (SBA) described in Section 5.2.6 was proposed and used in the 7/SBC/AQF coder.

The IBAP pattern was obtained using the two sentences of male and female speech data files described in Section 3.7, as training sequence. The noise shaping parameter \( \gamma \) was set to \(-0.3\) and the maximum number of bits allowed was clamped to 5. Unlike the other previously described 7/SBC coders which operate at 34.2 kbps, the total number of bits in the IBAP pattern was kept to 15 so that the
coder bit rate was close to the required 32 kbps. With side-information included, the total bit rate was 32.2 kbps. The IBAP pattern obtained from the training sequence was found to be

\[(1 \times 5 \text{ bits}, 2 \times 3 \text{ bits}, 2 \times 2 \text{ bits}, 2 \times 0 \text{ bit}).\]

To make up for the slight reduction in the number of bits available for the quantization of the subband signals, the formerly used AQF with uniform step-size was replaced by the one whose input-output characteristic was designed for signals which had a Laplacian distribution in amplitude. The detailed design of the Laplacian AQF is given in Appendix IV(b).

5.3.4.1 Results and Discussions

For comparison purposes, the 7/SBC/SBA scheme was simulated using speech data files which were used as a training sequence to derive the IBAP pattern, and also speech data files which were not used as a training sequence. The speech data files that were not used to form the training sequence were the ones described in Section 3.6. The average SNRSEG measurements for both sets of speech data files and for both the 7/SBC/ABA and 7/SBC/SBA coders are given in Tables 5.3.10 and 5.3.11. The same IBAP pattern was used for both sets of data files. Thus, it was a test of whether the IBAP was source dependent.
As can be seen from the average SNRSEG measurements, the use of the SBA algorithm only caused a drop of 0.3 and 0.5 dB compared to the use of the ABA algorithm for both sets of speech data files. Also, the IBAP pattern did not seem to be speaker or sentence dependent. Subjective listening tests revealed that there was virtually no difference in perceptual quality for the recovered speech for the 7/SBC coder employing either the ABA or SBA algorithm. This suggests that the use of the complicated ABA algorithm is truly unwarranted. The highly simple SBA algorithm can achieve as good performance for the coder as the ABA algorithm at this bit rate without incurring any additional side-information or system complexity. Its effectiveness will be further examined in the studies of the fourteen-band SBC coder and the large block-size ATC coder in the subsequent section and chapter.
Table 5.3.10  SNRSEG performance of the 7/SBC/ABA \((\gamma = -0.3)\) and 7/SBC/SBA coders. The sentences were used as a training sequence to obtain the IBAP pattern.

<table>
<thead>
<tr>
<th>Speech sentence</th>
<th>7/SBC/ABA</th>
<th>7/SBC/SBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>18.61</td>
<td>18.41</td>
</tr>
<tr>
<td>Female</td>
<td>19.92</td>
<td>19.57</td>
</tr>
<tr>
<td>Average</td>
<td>19.27</td>
<td>19.00</td>
</tr>
</tbody>
</table>

Table 5.3.11  SNRSEG performance of the 7/SBC/ABA \((\gamma = -0.3)\) and 7/SBC/SBA coders. The sentences were NOT used as the training sequence for the determination of the IBAP pattern.

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
<th>7/SBC/ABA</th>
<th>7/SBC/SBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.75</td>
<td>23.45</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>22.33</td>
<td>22.09</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>21.64</td>
<td>21.04</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>19.95</td>
<td>19.08</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>21.92</td>
<td>21.41</td>
<td></td>
</tr>
</tbody>
</table>
5.3.5 Seven-Band Subband Coding with Adaptive Pre- and Post-
Filtering, Fixed Bit Allocation and AQFs

The use of adaptive pre- and post-filtering (PPF) was found to be effective in the last chapter in achieving spectral shaping of the output noise in ADPCM coding of speech. Similarly, it can be incorporated into a subband coding scheme to provide another alternative method for achieving noise spectral shaping without using the adaptive bit allocation algorithm. The subband coder employing the PPF technique has the same structure as that of the 7/SBC/FBA coder described in Section 5.3.1 with the addition of an adaptive pre-filter before the subband encoder at the transmitter and the corresponding post-filter after the subband decoder at the receiver. The pre-filter is an adaptive 4th order pole-zero filter with its coefficients updated at every 16 msec. Its transfer function is given by

\[
1 - R(Z) = \frac{1 - A(Z)}{1 - A(Z/\beta)}
\]  

(5.3.3)

where \( \beta \) is the noise shaping parameter to be optimized using subjective listening tests and \( A(Z) \) is the conventional spectral envelope predictor. The post-filter is the exact inverse of the pre-filter.

As the transmission of the four pre-filter coefficients required 1.5 kbps, the total bit rate of the system is now equal to 35.7 kbps.
The number of bits allocated to the seven bands are fixed to (4, 3, 3, 2, 2, 1, 1) for the 1st to the 7th band. Uniform AQFs are employed for the quantization of the subband signals.

5.3.5.1 Results and Discussions

To obtain the best perceptual performance for the 7/SBC coder employing the PPF noise shaping technique, the value of the shaping parameter $\beta$ of the pre-filter was found to be 0.7. The SNRSEG measurements of the coder for the four different speech sentences are given in Table 5.3.12. The values are lower than that of the scheme employing the ABA technique of noise shaping. The noise spectral plots of the output noise of the coder in comparison with that of the 7/SBC/ABA scheme are given in Figure 5.3.9. Adaptive pre- and post-filtering clearly achieves finer spectral shaping of the quantization noise compared with the ABA technique. In the latter, the noise level is flat in each band and the bit assignment rule given by Equation (5.2.29) can only achieve coarse spectral shaping. Subjective listening tests did not reveal any significant difference between them, suggesting that the use of the complicated pre- and post-filtering is not really justified. Furthermore, additional side-information about the filter's coefficients has to be transmitted, making it less attractive as a possible candidate for the coding of wideband speech at 2 bits/sample.
### Table 5.3.12  SNRSEG measurements of the 7/SBC/PPF/FBA scheme with $\beta = 0.7$

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20.25</td>
</tr>
<tr>
<td>2</td>
<td>18.35</td>
</tr>
<tr>
<td>3</td>
<td>16.51</td>
</tr>
<tr>
<td>4</td>
<td>16.69</td>
</tr>
<tr>
<td><strong>Average</strong></td>
<td><strong>17.95</strong></td>
</tr>
</tbody>
</table>

### Figure 5.3.9  Long-term average power spectral density plots of

1. Original speech
2. Output noise of the 7/SBC/ABA system with $\gamma = -0.3$
3. Output noise of the 7/SBC/PPF/FBA system with $\beta = 0.7$
5.3.6 Summary of the Performances of the 7/SBC Coders

The investigation of 7/SBC coders have shown that it is possible to obtain satisfactory quality wideband speech at a bit rate of slightly greater than 32 kbps, using the various bit allocation strategies or the adaptive pre- and post-filtering techniques. The 7/SBC/ABA and 7/SBC/PPF/FBA schemes were found to produce better perceptual performance compared to the schemes employing the FBA strategy. For the 7/SBC/FBA schemes, AQF seemed to be a better quantization technique than ADPCM/AQJ.

The proposed simplified bit allocation algorithm had been proved to be a valuable alternative to the full ABA algorithm. There was no perceptual difference for the 7/SBC coder employing either the ABA or SBA algorithm.

In comparison with the ADPCM/ANS coding techniques investigated in the previous chapter, 7/SBC coders are undoubtedly superior even for the schemes with fixed bit allocation. The confinements of noise in each individual bands and the masking properties of speech are more effectively exploited to produce better quality speech in subband coding. However, the overall quality, even for the scheme employing the ABA or SBA algorithm, is still slightly below perfect for the truly good quality wideband services application. To further improve the subjective quality of the recovered speech, one possible solution is to increase the number of bands from 7 to 14. Studies of the various 14-band SBC coders are presented in the following section.
The performance of a subband coder can be improved to a certain extent by increasing the number of subbands to improve the frequency resolution of the coder. A higher number of subbands will give rise to more flexibility in the use of the ABA algorithm and result in finer spectral control of the output noise. However, there is a limit to the number of subbands a coder can have. A higher number of subbands implies greater system complexity and longer coder delay. In the existing literature, the highest number of subbands ever used in the design of a subband coding scheme is 32\(^{(79)}\). Most SBC coders employ 16 or less subbands.

In the design of fourteen-band subband coders (14/SBC) for the coding of wideband speech, 32-tap FIR filters\(^{(81)}\) are used throughout the four-stage tree-structured analysis and synthesis filter-banks. The total coder delay due to the filter-banks can be calculated using Equation (5.3.1) to be 465 samples which is equivalent to 29 msec. It can be reduced by using lower order FIR filters for the higher stages of the filter bank. If AQFs are used for the quantization of the subband signals, an additional delay of 16 msec is introduced by the buffering of 16 msec of subband signals for the calculation of the AQF step-sizes. The side-information used for the transmission of the step-sizes of the 14-band system is increased to 4.4 kbps which is double that of the 7-band scheme.
More bits used for the transmission of side-information means less bits are available for the quantization of the subband signals, if the overall system bit rate is to be maintained at constant. Alternatively, AQJ can be employed for the quantization of the subband signals and side-information can be cut down to the minimum.

Two fourteen-band subband coders with different bit allocation and quantization strategies were examined for the coding of wideband speech at 32 kbps. They are

(1) 14/SBC coder with fixed bit allocation and AQJs for the first 12 bands and AQFs for the last two bands.

(2) 14/SBC coder with the ABA and SBA algorithms. Laplacian AQFs are employed for the quantization of the subband signals.

Detailed description of the coders and their performances are presented in the following sections. Transmission errors were introduced also in the simulation of the operation of the 14/SBC coder employing fixed bit allocation.
5.4.1 Fourteen-Band Subband Coding (14/SBC) with Fixed Bit Allocation

ADPCM/AQJ and AQF Encoders

The system structure of the 14/SBC coder is similar to that of the 7/SBC coder shown in Figure 5.2.1. The tree-structured QMF analysis and synthesis filter-banks are extended to four stages using the 32-tap FIR filter designed by Johnston. For the first simulation of the 14/SBC scheme, fixed bit allocation was employed. The first twelve bands employed ADPCM/AQJ encoders and therefore no side-information is required. The highest two bands use one-bit AQF quantizers and require 625 bps for the transmission of the step-sizes as side-information. System configuration of the coder is summarized in Table 5.4.1. The ADPCM encoders for the first twelve bands employ fixed first order predictors and the leakage constant of all the AQJ was set to $\beta = 1 - 2^{-5}$. The operations of the coder were simulated both with and without transmission errors. Two sets of speech data were used as input to the coder. The first set consists of the four sentences described in Section 3.6 and the second set consists of the two sentences described in Section 3.7.

5.4.1.1 Results and Discussions

(a) First set of data:

The average SNRSEG measurement of the 14/SBC coder employing fixed bit allocation and ADPCM/AQJ was found to be 15.70 dB compared to 25.43 dB for the 7/SBC/ABA scheme.
<table>
<thead>
<tr>
<th>Band No.</th>
<th>Bit Allocation (fixed)</th>
<th>Encoder</th>
<th>Predictor Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>ADPCM/AQJ</td>
<td>-0.2178</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>&quot;</td>
<td>-0.6156</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>&quot;</td>
<td>0.2689</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>&quot;</td>
<td>-0.1104</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>&quot;</td>
<td>0.1275</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>&quot;</td>
<td>0.1438</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>&quot;</td>
<td>0.3373</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>&quot;</td>
<td>0.0279</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>&quot;</td>
<td>0.1587</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>&quot;</td>
<td>-0.1317</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>&quot;</td>
<td>-0.0254</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>&quot;</td>
<td>-0.1725</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>AQF</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>&quot;</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.4.1  System description of the 14-band subband coder with fixed bit allocation
It should be noted that the 7/SBC coder operates at 34.2 kbps whereas the 14/SBC coder operates at 31.63 dB. The long-term average power spectral density plots of the output noise of the two coders are shown in Figure 5.4.1. Finer spectral control of the noise of the 14/SBC coder can be observed. However, the overall noise level is too high, especially for the first two bands, to be effectively masked by the speech signals. To reduce the level of noise in the first two bands, one bit each may be taken from the 11th and 12th and re-allocated to the first two bands. However, this will incur another 625 bps of side-information for the transmission of the step-sizes of the one-bit AQF of the 11th and 12th band. The system bit rate will then be 32.25 kbps. Subjectively, the perceptual difference between the recovered speech of the 14/SBC/FBA coder and that of the 7/SBC/FBA coder was not as great at that suggested by the big difference of 10 dB in SNRSEG measurement. The overall speech quality of the 14/SBC coder is still not satisfactory.

(b) Second set of data:

The same simulation was carried out for the second set of input data both with and without transmission error. With no transmission error, the average SNRSEG measurement of the 14/SBC coder was found to be 13.46 dB (see Table 5.4.3). At the bit error rates of $10^{-4}$ and $10^{-3}$, SNRSEG decreases by only 2.5 and 3.6 dB respectively. This compares well with the two 64 kbps schemes (2-band SBC and ADPCM) which suffer SNRSEG reduction of around 10 dB at the BER of $10^{-3}$. The robustness of the coder under transmission error conditions is due to:
Table 5.4.2 SNRSEG measurements for the 14/SBC coder with FBA and ADPCM/AQJ encoders

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>SNRSEG (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>16.31</td>
</tr>
<tr>
<td>2</td>
<td>17.85</td>
</tr>
<tr>
<td>3</td>
<td>13.98</td>
</tr>
<tr>
<td>4</td>
<td>14.67</td>
</tr>
<tr>
<td>Average</td>
<td>15.70</td>
</tr>
</tbody>
</table>

Figure 5.4.1 Long-term average power spectral density plots of

(1) Original speech
(2) Output noise of the 7/SBC/ABA coder
(3) Output noise of the 14/SBC/ADPCM/AQJ coder
Table 5.4.3  SNRSEG measurements for the 14/SBC coder with ADPCM/AQJ encoders under the different BER conditions

<table>
<thead>
<tr>
<th>Sentence</th>
<th>BER 0</th>
<th>BER $10^{-4}$</th>
<th>BER $10^{-3}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>12.18</td>
<td>9.78</td>
<td>8.50</td>
</tr>
<tr>
<td>Female</td>
<td>14.73</td>
<td>12.41</td>
<td>11.65</td>
</tr>
<tr>
<td>Average</td>
<td>13.46</td>
<td>11.10</td>
<td>10.08</td>
</tr>
</tbody>
</table>

Figure 5.4.2  SNRSEG performance of the 14/SBC coder against BER
(1) The twelve AQJs used provided more dissipation power for transmission errors which might occur in any band.

(2) The effect of transmission errors is largely confined to each band due to the filtering operation of the synthesis filter-bank at the receiver.

The subjective quality of the recovered speech of the coder was found to be slightly below satisfactory especially when listened through head-phones. Adaptive bit allocation and the use of AQFs are necessary to further improve the quality of the recovered speech of the coder.
5.4.2 Fourteen-Band Subband Coding (14/SBC) Employing the Full ABA or SBA Algorithm and Laplacian AQF Encoders

The second 14-band subband coding scheme employs Laplacian AQFs for the quantization of all the subband signals. The step-sizes of the AQFs are updated every 16 msec and transmitted to the receiver as side-information, this required a total bit rate of 4.4 kbps and the coder is therefore left with 27.6 kbps for the encoding of the subband signals. Laplacian AQF gives better performance than uniform AQF because the pdf of the subband signals is known to be closer to a Laplacian pdf than to a uniform pdf. The input threshold levels and the output values of a Laplacian quantizer for different number of bits are given in Appendix IV(b).

Two adaptive bit allocation algorithms are employed. The first one is the ABA algorithm given by Equation (5.2.29) with $\gamma = -0.3$ and the maximum number of bits restricted to 5. The second algorithm is the SBA algorithm described in Section 5.2.6. Three IBAP patterns for the 14/SBC/AQF coder for the transmission bit rates of 32, 24 and 16 kbps were obtained using the two male and female sentences described in Section 3.7 as training sequence. They are given in Table 5.4.4.

<table>
<thead>
<tr>
<th>Bit rate (kbps)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>24</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 5.4.4 The three time-variant bit allocation patterns (IBAP) for the 14/SBC coder for the transmission bit rates of 32, 24 and 16 kbps.
5.4.2.1 Results and Discussions

The SNRSEG measurements of the 14/SBC/AQF/ABA coder are given in Table 5.4.5 and that of the same coder employing the SBA algorithm are given in Table 5.4.6. There is only a drop of 0.6 dB for using the SBA instead of the ABA algorithm at 32 kbps. Virtually no difference in the perceptual quality of the recovered speech could be found between the two coders. As the sentences used in the simulations are the ones used as the training sequence to derive the IBAP patterns, one may suspect the effectiveness of the SBA algorithm for speech sentences which are out of the training sequence. To test whether the IBAP patterns are source dependent, the same IBAP pattern for 32 kbps was used in the coding of the other four different sentences described in Section 3.6. The SNRSEG performance of the 14/SBC/AQF coder employing the ABA or SBA algorithm for the four speech data files which are out of the training sequence are given in Table 5.4.7. Again, there is only a drop of 0.75 dB in average SNRSEG performance when using the SBA instead of the ABA algorithm for this set of data files. This clearly suggests that the use of the SBA algorithm is truly a valuable substitute for the ABA algorithm.

At the bit rate of 32 kbps, both the 14/SBC/ABA and 14/SBC/SBA coders produce recovered speech of near excellent quality. When the bit rate is reduced to 24 kbps, the output noise is just at the threshold of audibility. However, the overall subjective quality of the recovered
<table>
<thead>
<tr>
<th>Sentence</th>
<th>Transmission Bit Rate (kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>19.26</td>
</tr>
<tr>
<td>Female</td>
<td>20.52</td>
</tr>
<tr>
<td>Average</td>
<td>19.88</td>
</tr>
</tbody>
</table>

Table 5.4.5  SNRSEG performance (in dB) of the 14/SBC/AQF/ABA coder at 3 different transmission bit rates

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Transmission Bit Rate (kbps)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>32</td>
</tr>
<tr>
<td>Male</td>
<td>18.79</td>
</tr>
<tr>
<td>Female</td>
<td>19.68</td>
</tr>
<tr>
<td>Average</td>
<td>19.24</td>
</tr>
</tbody>
</table>

Table 5.4.6  SNRSEG performance (in dB) of the 14/SBC/AQF/SBA coder at 3 different transmission bit rates

<table>
<thead>
<tr>
<th>Sentence No.</th>
<th>14/SBC/AQF coder employing ABA</th>
<th>SBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.83</td>
<td>22.90</td>
</tr>
<tr>
<td>2</td>
<td>22.64</td>
<td>21.97</td>
</tr>
<tr>
<td>3</td>
<td>21.88</td>
<td>20.91</td>
</tr>
<tr>
<td>4</td>
<td>19.30</td>
<td>18.84</td>
</tr>
<tr>
<td>Average</td>
<td>21.91</td>
<td>21.16</td>
</tr>
</tbody>
</table>

Table 5.4.7  SNRSEG performance (in dB) of the 14/SBC/AQF coder employing the ABA and SBA algorithms.

The sentences were not used as the training sequence to derive the IBAP pattern for the SBA algorithm.
speech could still be considered as satisfactory. The slight difference in the perceptual quality of the recovered speech of the 14/SBC/AQF/SBA coder when operating at the different bit rates of 32 and 24 kbps suggests that it is a very suitable choice as a variable bit rate speech transmission system. Also, the use of the SBA algorithm facilitates the switching of the operating bit rate of the coder.

At the bit rate of 16 kbps, the output noise consists of both the quantization noise and the aliasing noise. The latter is due to the fact that poor quantization or zero bit allocation for the quantization of the subband signals does not ensure that aliasing noise is perfectly cancelled off at the receiver. Though with more distortions, the recovered speech at this bit rate did not sound too unpleasant due to the wider bandwidth, i.e. 7 kHz instead of 3.4 kHz, of the system. Its performance in comparison with the 14/SBC/ABA coding of narrowband (0 - 3.4 kHz) speech at the same bit rate will be discussed in the next section.

At the bit rate of 32 kbps, the average SNRSEG measurement of the 14/SBC/AQF/SBA coder has a value of 19.24 dB compared to 13.68 dB of the 14/SBC/ADPCM/FBA coder. The comparison of the long-term average power spectral density plots of the two coders are given in Figure 5.4.3. These two objective measurements suggest that the coder with AQF/SBA has a clear lead over the one with ADPCM/FBA. The reduction of the noise power by 5.6 dB when going from ADPCM/FBA to AQF/SBA together
with the maintenance of fine spectral control of the output noise lead
to considerable improvement in the perceptual quality of the recovered
speech. This suggests that it is more advantageous to use AQF and SBA
rather than ADPCM/AQJ and FBA.

The comparison between the average SNRSEG measurements of 14/SBC/SBA
coder and the 7/SBC/SBA coder only shows a difference of 0.88 dB with
the former having a higher value of 19.88 dB. However, the 14/SBC
coder was found to outperform the 7/SBC coder in terms of subjective
listening tests due to the finer control of the output noise arising
from the larger number of subbands of the former. The spectral
distribution of the power of the output noise of the two coders are shown
in Figure 5.4.4.
Figure 5.4.3 Long-term average power spectral density plots of
(1) Original Speech
(2) Output noise of the 14/SBC/FBA/ADPCM coder
(3) Output noise of the 14/SBC/SBA/AQF coder

Figure 5.4.4 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the 7/SBC/SBA/AQF coder
(3) Output noise of the 14/SBC/SBA/AQF coder
5.4.3 Comparison between the 14/SBC/AQF Coding of Wideband and Narrowband Speech at the Bit Rate of 16 kbps

When the bit rate for the coding of wideband speech is reduced from 32 to 16 kbps, there is only an average of 1 bit/sample for the encoder to describe the speech waveforms in the best possible way. Signal distortion is expected to be high in the case of waveform coding. The usual practice is to employ vocoding techniques or the various forms of frequency compression schemes followed by frequency-domain waveform coders at this bit rate. The 14/SBC/AQF coding of wideband speech at 16 kbps was not unexpectedly found to be corrupted by a considerable amount of quantization and aliasing noise especially during voiced speech. However, there is an important difference between the 16 kbps wideband and narrowband speech coders. Though the former is very clearly corrupted by noise, it possesses a much wider bandwidth than the latter. For comparison purposes, a 14/SBC/AQF/ABA coding of narrowband (0 - 3.4 kHz) speech was simulated. The coder has an average SNRSEG measurement of 17.74 dB as shown in Table 5.4.8.

Informal subjective listening tests showed that the quality of the recovered speech of the narrowband coder was as good as that of the unprocessed narrowband speech. However, as both the original and coded narrowband speech are bandlimited to 3.4 kHz, the intelligibility of unvoiced sounds was reduced. Unlike the narrowband speech, the wideband speech coder does not seem to have intelligibility problems due to the wider bandwidth it has. This suggests that at the bit rate of 16 kbps,
one has the choice of whether to have toll quality but bandlimited speech or wideband speech but with more quantization and aliasing distortion. The former has the problem of intelligibility for unvoiced sounds and the latter is degraded by a considerable amount of noise.

<table>
<thead>
<tr>
<th>Sentence</th>
<th>SNRSEG (in dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>16.50</td>
</tr>
<tr>
<td>Female</td>
<td>18.98</td>
</tr>
<tr>
<td>Average</td>
<td>17.74</td>
</tr>
</tbody>
</table>

Table 5.4.8. The average SNRSEG performance of a 14/SBC/AQF/ABA narrowband speech coder operating at 16 kbps.
Comparison Between the 14/SBC/AQF Coder at 32 kbps and the 2-band SBC Coder at 64 kbps

One of the objectives for the design of a wideband speech coder at 32 kbps is to achieve as good quality as that of the 64 kbps coder for the recovered speech. The investigation of 32 kbps ADPCM coding of wideband speech employing the technique of adaptive noise spectral shaping described in Chapter 3, shows that the quality of the recovered speech, using this kind of time-domain waveform coding technique, is undoubtedly below that of the 64 kbps 2-band SBC scheme. The more promising subband coding technique employing fixed bit allocation and AQJs, described in the previous sections of this chapter, is degraded also by a low but audible level of noise and this is no match with the performance of the 64 kbps coder. However, the 14-band SBC coder employing the fully adaptive bit allocation or the simplified bit allocation algorithm produces near excellent subjective quality for the recovered speech. Informal subjective listening tests were carried out to compare the 14/SBC/AQF/SBA system at 32 kbps with the 2-band SBC system at 64 kbps.

The long-term average spectral plots of the output noise of the two systems are given also in Figure 5.4.5. Though the 32 kbps 14/SBC coder has a higher level of noise than the 64 kbps 2-band SBC coder, its spectral control of the noise is finer. Perceptually, the recovered speech of both coders are almost as good as the original unprocessed
wideband speech. It was found almost impossible to distinguish the recovered speech of the two coders despite the fact that they have an SNR difference of 5.8 dB. Thus, SNR measurements and average spectral plots of the output noise can only serve as complementary performance 'yardsticks' and design aids.

Figure 5.4.5 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the 14/SBC/AQF/SBA coder operating at 32 kbps
(3) Output noise of the 2-band SBC coder operating at 64 kbps
5.5 **SEVEN-BAND COMPLEX SUBBAND CODERS (7/CSBC)**

This section presents the use of the relatively unknown complex quadrature mirror filter (CQMF) in the design of seven-band complex subband coders (7/CSBC) for the coding of wideband speech at 32 kbps. The design concept of quadrature mirror filter can be extended to the design of complex quadrature mirror filter which decomposes a signal into real and imaginary components and provides aliasing free reconstruction. The real and imaginary components at the output of a CQMF can be used to form the amplitude and phase signals which provide an alternative representation of the signal. Conventional subband coders employ QMF filter banks for the division of full band speech signals into a number of different frequency subbands. As described in Section 5.2.4, if CQMF, are connected to the last stage of the tree-structured QMF filter bank, a complex QMF filter bank is obtained. The amplitude and phase signal of each subband of the coder can then be encoded using different algorithms.

The decomposition of a signal into amplitude and phase using the DFT is well known and sometimes used for the design of speech coder (167). However, a subband coder with its design based on CQMF to achieve similar decomposition function and its performance for the encoding of wideband or narrowband speech are relatively new areas of investigation. For the coding of wideband speech at 32 kbps, two seven-band CSBC coders employing different bit allocation and quantization strategies are proposed. They are
(1) 7/CSBC coder employing the novel simplified bit allocation (SBA) algorithm and Laplacian AQF for the quantization of the amplitude signals and uniform AQF for the phase signals.

(2) 7/CSBC coder employing seven fixed bit allocation patterns and a novel quantization technique, which exploits the correlation between amplitude signals from the adjacent bands.
5.5.1 Seven-Band Complex Subband Coding (7/CSBC) with simplified Bit Allocation and AQFs

The basic system block diagram of a seven-band complex subband coder is shown in Figure 5.5.1. The 32-tap FIR lowpass filter described in Section 5.3.1 is used to form the CQMF analysis and synthesis filter banks. The total delay due to the filter banks is therefore the same as that of the 14/SBC scheme described previously. The CQMF analysis filter bank at the transmitter divides the full-band speech signal into seven uniform pairs of complex channels. The real \( R \) and imaginary \( I \) signals of each pair of complex channels are used to form the amplitude \( A \) and phase \( \phi \) signals which are quantized using Laplacian and uniform AQFs respectively.

The RMS value \( \sigma_i \) of a block of 16 msec (16 samples) of the amplitude signal, \( A_i \), of the ith band is computed and quantized with a 5-bit Gaussian AQF and transmitted to the receiver. (The characteristics of Gaussian quantizer are given in Appendix IV). As there are seven \( \sigma_i \)'s, the total bit rate for the transmission of \( \sigma \) is 2.2 kbps. The coder is therefore left with 30 kbps for the quantization of \( A \) and \( \phi \). The value \( \hat{\sigma}_i \)'s (quantized value of \( \sigma_i \)'s) are used to scale the amplitude signal \( A_i \)'s which are then quantized by a unit variance Laplacian quantizer.

The technique of simplified bit allocation algorithm described in Section 5.2.6 is employed for the coding of the amplitude and phase
Input Speech S
16 kHz

CQMF Analysis Bank

Rectangular to Polar Conversion

Power Estimation

Multi-level to Binary

Laplacian Quantizer

Derive Bit Allocation

Uniform Quantizer

B(\hat{\theta})

To channel

(a) Transmitter

Note: B(\hat{x}) means quantized value of \hat{x} in binary

(b) Receiver

Figure 5.5.1 System block diagram of a complex subband coder
signals. The IBAP pattern used for the 32 kbps is (5, 4, 2, 2, 2, 0, 0). The subbands are ranked in descending order according to the magnitudes of their step-sizes, $\sigma_i$'s, which are known both at the receiver and transmitter. The first band with the highest step-size is allocated 5 bits, the second higher energy band with 4 bits and so on. This procedure is repeated every 16 msec. The IBAP patterns of the SBA algorithm for the quantization of the amplitude and phase signals at three different bit rates of 32, 24 and 16 kbps are given in Table 5.5.1. When 0 bit is allocated to a subband amplitude channel, the amplitude information of the most significant band, i.e. the band having the largest step-size, is copied to that band and scaled down according to their step-size ratio. This direct duplication procedure is adopted based on the observation that the amplitude signals of all the subbands have similar temporal variations for voiced speech as shown in Figures 5.5.2a and b. Coarse description of the amplitude information by using the duplication procedure is acceptable for the bands with 0 bit allocation because after all, they are perceptually less significant and the ear can tolerate a certain degree of distortion in the recovered speech.

Uniform quantization of the phase signals is employed because they exhibit uniform distribution between the values of $\pm \pi$. The angle from $+\pi$ to $-\pi$ is uniformly divided into $2^B$ uniform angular regions where $B$ is the number of bits allocated. For instance, when $B$ is equal to 3, the four
regions in the positive side will be \( 0 + \frac{\pi}{4}, \frac{\pi}{4} + \frac{\pi}{2}, \frac{\pi}{2} + \frac{3\pi}{4} \) and \( \frac{3\pi}{4} + \pi \).

If the input phase falls within the \( 0 + \frac{\pi}{4} \) region, it would be quantized to \( \frac{\pi}{8} \) and so on. When 0 bits are allocated to a subband phase channel, a random phase is generated at the receiver for that band.

<table>
<thead>
<tr>
<th>Bit rate (kbps)</th>
<th>IBAP Patterns</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Amplitude phase 5 4 2 2 2 0 0 0</td>
<td>30</td>
</tr>
<tr>
<td>24</td>
<td>Amplitude phase 5 3 2 2 2 0 0 0</td>
<td>22</td>
</tr>
<tr>
<td>16</td>
<td>Amplitude phase 4 2 2 0 0 0 0 0</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 5.5.1 The time-variant bit allocation patterns of the simplified dynamic bit allocation algorithm for 3 different transmission bit rates.
Figure 5.5.2a Amplitude signals of the subbands of a 7-band CSBC system for voiced speech
Figure 5.5.2b  Normalized amplitude signals of the subbands of a 7-band CSBC system for voiced speech
Figure 5.5.2c Amplitude signals of the subbands of a 7-band CSBC system for voiced and unvoiced speech
5.5.1.1 Results and Discussions

The operation of the 7/CSBC coder employing the simplified bit allocation algorithm and AQFs was simulated at the three transmission bit rates of 32, 24 and 16 kbps. The average SNRSEG measurements of the coder are given in Table 5.5.2. At 32 kbps, the average SNRSEG of the coder was found to be 18.63 dB compared to 19.24 dB of the 14/SBC/AQF/SBA coder and 19.00 of the 7/SBC/AQF/SBA coder. Comparison in terms of long-term average power spectral density plots of the output noise of the 7/CSBC and 14/SBC coders are given in Figure 5.5.3. It is obvious that the output noise in the lower bands of the 7/CSBC coder is lower than that of the 14/SBC coder. However, the 7/CSBC coder has a higher level of noise in the upper bands than the 14/SBC coder. Subjective listening tests showed that only very slight degradation in the quality of the recovered speech of the 7/CSBC coder can be heard. Therefore it is only marginally worse than the excellent 14/SBC coder. Subjective listening tests revealed also that it was almost impossible to perceive any difference between the recovered speech of the 7/SBC and 7/CSBC coders. This is due to the same frequency resolution and thus the same degree of control of the output noise of the two coders. The overall quality of the 7/CSBC coder can be described as good.

At the lower bit rates of 24 and 16 kbps, the 7/CSBC coder produces recovered speech of slightly worse perceptual quality than that of the recovered speech of the 14/SBC coder operating at the same bit rates.
Table 5.5.2 SNRSEG performance of the seven-band
CSBC/AQF/SBA coder at 3 transmission
bit rates

<table>
<thead>
<tr>
<th>Transmission Bit Rate (kbps)</th>
<th>32</th>
<th>24</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>17.82</td>
<td>15.80</td>
<td>11.48</td>
</tr>
<tr>
<td>Female</td>
<td>19.43</td>
<td>17.96</td>
<td>13.54</td>
</tr>
<tr>
<td>Average</td>
<td>18.63</td>
<td>16.88</td>
<td>12.51</td>
</tr>
</tbody>
</table>

Figure 5.5.3 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the 14/SBC/AQF/SBA coder
(3) Output noise of the 7/CSBC/AQF/SBA coder at 32 kbps
though their average SNRSEG measurements are almost the same. The subjective quality at 24 kbps is below but near satisfactory. At 16 kbps, the recovered speech is degraded by large amounts of quantization and aliasing noise. Subjectively, it is unacceptable.

5.5.2 Seven-Band Complex Subband Coding (7/CSBC) with Seven Fixed Bit Allocation Patterns and a Novel Quantization Technique

The system design of the second 7/CSBC scheme is similar to the one described in the previous section except the following two main differences:

(1) Instead of using the simplified bit allocation algorithm, the number of possible bit allocation patterns is restricted to seven. The determination of the seven bit-allocation patterns is described as follows:

In a training sequence of voiced and unvoiced sounds, a normalized vector \( \hat{A} \) (\( \sigma_1^2 / \sigma_T^2 \), \( \sigma_2^2 / \sigma_T^2 \), ..., \( \sigma_7^2 / \sigma_T^2 \)) (where \( \sigma_T^2 = \sum_{i=1}^{7} \sigma_i^2 \)) is obtained every 16 msec from the amplitude signals \( A_i \) of the 7 channels. \( \sigma_i^2 \) is the mean squared value of \( A_i \)'s of the ith-band within 16 msec. Normalization is necessary as the input speech power varies from frame to frame. A collection of \( \hat{A} \)'s are then vector quantized to 8 vectors \( V_i \)'s (\( i = 1, \ldots, 8 \)), which are then used to obtain the 8 bit allocation patterns using the ABA algorithm given by Equation (5.2.29) with
\[ \gamma = -0.3. \] The 8 bit allocation patterns are given in Table 5.5.3. The technique of vector quantization of multi-dimensional vectors can be found in References 164 and 165. In the process of vector quantization, the well known minimum Euclidean distance criterion was employed. The first 4 bit allocation patterns in Table 5.5.3 clearly corresponds to voiced speech and the last 4 are undoubtedly arising from fricative and plosive sounds. As the maximum of 5 bits is usually sufficient for the quantization of subband signals, the first two patterns are combined to one as shown in the table and the total number of patterns is reduced to seven.

The bit allocation patterns for the quantization of the phase signals are obtained by modifying the seven bit allocation patterns for the amplitude signals so that the total bit rate of the coder is about 32 kbps. The seven patterns for the phase corresponding to the seven patterns for the amplitude, are given in Table 5.5.4.

As the vector A is transmitted to the receiver every 16 msec, it can therefore be used to decide which bit allocation pattern to use by simply measuring the 8 Euclidean distances \[ d_i = (V_i - A) \cdot (V_i - A)^T \] for \( i = 1, \ldots, 8 \). This particular form of bit allocation technique will be represented by VBA hereafter.
Table 5.5.3  The bit allocation patterns for the quantization of the amplitude signals

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>combined to</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>} 3 3 2 1 0 0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.5.4  The bit allocation patterns for the quantization of the phase signals

<table>
<thead>
<tr>
<th>Pattern No.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>
Once the bit allocation pattern is determined, the next step is to encode the amplitude and phase signal for transmission. Instead of employing the conventional AQF encoders, different strategies are proposed for the quantization of the amplitude and phase signals.

**Amplitude signals:**
To achieve more accurate quantization, the temporal correlation between the samples within each band, or the inter-band correlation between the samples from the adjacent bands must be exploited. In the temporal domain, the periodic structure of the amplitude waveforms (see Figure 5.5.2.b), arising from the pitch structure of the voiced speech, suggests that pitch prediction can be applied in a differential system to encode the amplitude signals. However, it will mean a substantial increase in system complexity as pitch predictor is required for each band and side-information pertaining to the pitch periods has to be transmitted. Instead, the inter-band correlation between the samples of the adjacent bands is exploited to achieve improved quantization performance.

The amplitude waveforms shown in Figure 5.5.2.b shows that during voiced speech, small amplitude samples in one band invariably correspond to small amplitude samples in the adjacent bands and similarly, large amplitude samples in one band also correspond very likely to large amplitude samples in the adjacent bands.
This observation suggests the following quantization procedure:

Take the first bit allocation pattern (5 3 3 2 1 0 0) for example. The 5th-band is allocated 1 bit which corresponds to 2 quantization output levels '0' and '1'. The histogram of all the amplitude samples in this band, when this bit allocation pattern is chosen, is obtained and used to design an optimum 1 bit quantizer using the technique of vector quantization in the one-dimension (164). For '0' output in this band, the corresponding amplitude samples in the 4th-band are collected and their histogram plotted (Figure 5.5.4a). Similarly, for '1' output in this band, the corresponding amplitude samples in the 4th-band are collected and their histogram plotted (Figure 5.5.4.b).

The first amplitude histogram has a relatively peaky distribution and smaller value of mean than that of the second. Two optimum 2-bit quantizers for the 4th-band, Q1 and Q2, are designed to match these two different amplitude histograms for improved SNR performance. There are 4 possible output levels for both Q1 and Q2. For the amplitude samples which give rise to one of the two inner output levels of Q1 and Q2, the corresponding samples in the 3rd-band, most likely to be small in amplitude, are collected and their histogram plotted (figure 5.5.5a).

Similarly for the samples that give rise to one of the two outer levels of Q1 or Q2, the corresponding samples in the 3rd-band are collected and their histogram plotted (Figure 5.5.5b). The two amplitude histograms can be used to design two optimum quantizer for the 3rd-band,
with one of them for the predominantly large amplitude samples and the other for the small amplitude samples.

The drawback of this strategy is the large amount of Read-Only-Memory (ROM) required to store the input thresholds and output levels of all the quantizers. Take the case of 4-bit allocation for example, there are a total of four 4-bit allocations in two different bands and three different bit allocation patterns (see Table 5.5.2). Thus a total of 8 different quantizers are required. As all the histograms corresponding to the large/small amplitude samples are very similar regardless of which band and which bit allocation they are in, the 8 quantizers can be reduced to 2 with one of them designed to suit the small amplitude samples and the other for the large amplitude samples. The similar procedure is applied to other bit allocation and the resulting quantizer complexity is therefore only double that of the conventional approach using just one quantizer.
Figure 5.5.4 Histograms of the 4th-band amplitude samples which correspond to amplitude samples in the 5th band having (a) inner output level (b) outer output level during quantization
Figure 5.5.5  Histograms of the 3rd-band amplitude samples which correspond to amplitude samples in the 4th-band having (a) inner output level (b) outer output level during quantization.
Phase signals:

The straightforward way to quantize the phase signals is to use a uniform step-size quantizer and the same number of bits for each sample within every 16 msec interval. If the phase signal of a particular band is allocated 5 bits, there are a total of 80 bits \((= 5 \times 16)\) for the quantization of the 16 samples in that 16 msec. Out of the 16 phase samples, some of them correspond to large amplitude samples in the same band and the other correspond to small amplitude samples. To improve the overall SNR performance of the system, one can employ a very simple strategy of allocating more bits for the phase samples which correspond to large amplitude samples and less bits for the phase samples which correspond to small amplitude samples. As the locally decoded amplitude samples are available both at the transmitter and receiver, they can be used to determine the bit allocation patterns for the phase signals. The simplified bit allocation (SBA) algorithm described in Section 5.2.6 is employed in this case because it does not incur undue system complexity.

For the case of 5-bit allocation, there are a total of 80 bits for the 16 samples. The IBAP pattern is given below:

\[
5 \text{ bit: } \{5 \times 6, 7 \times 5, 3 \times 4, 1 \times 3\}.
\]

The first 5 phase samples which have the first 5 highest amplitude samples receive 6 bits each; the next 7 phase samples which have the most 7 highest amplitude samples receive 5 bits each and so on. The
other IBAP patterns are

4-bit : \{2 \times 5, 12 \times 4, 2 \times 3\},

3-bit : \{2 \times 4, 12 \times 3 2 \times 2\}.

For the case of 2-bit allocation, all 16 samples in that band receive the same number of 2 bits. During experimentation it was found that a minimum of 2 bits is required to represent the 4 quadrants of the phase signals. The quantization strategies of the amplitude and phase signals described so far will be denoted by CAQF/$\$SBA where C stands for correlation and $\$SBA means quantization of the phase signals with the SBA algorithm. The 7/CSBC coder employing the VBA algorithm and the CAQF/$\$SBA strategy will be denoted by 7/CSBC/CAQF/$\$SBA hereafter.

To test the effectiveness of the CAQF/$\$SBA technique, a 7/CSBC coder employing the VBA algorithm, the conventional Laplacian and uniform AQFs for the amplitude and phase, was simulated for comparison. This coder will be denoted by 7/CSBC/VBA/AQF.
5.5.2.1 Results and Discussions

The average SNRSEG measurements for the two 7/CSBC coders are given in Table 5.5.5. The use of the CAQF/SBA technique clearly leads to an SNR improvement of 1.5 dB compared to the conventional AQF quantization technique. The long-term average power spectral density plots of the output noise of the two coders shown in Figure 5.5.6 also indicate a lower level of noise power for the coder employing the CAQF/SBA technique. This improvement in SNR performance is obtained at a very slight increase in system complexity when going from AQF to CAQF/SBA. Also, no additional side-information is incurred by using the more efficient technique. Perceptually however, the difference in the quality of the recovered speech of the two 7/CSBC coders could hardly be perceived for the speech sentences tested. Nevertheless, if two coders possess the same coding characteristics, like frequency resolution and noise spectral control, and similar system complexity, the one with higher SNR performance is naturally preferred. Furthermore, the improvement in SNR performance might lead to an improvement in the subjective quality of the recovered speech for some other speech sentences spoken by other speakers.

The comparison in terms of the noise spectral plots of the 7/CSBC/SBA/AQF and 7/CSBC/VBA/AQF coders is shown in Figure 5.5.7. The coder employing the SBA algorithm has a lower level of output noise for the first band but at the expense of a higher level of noise in the third and the last
two bands compared to that of the coder employing the VBA algorithm. Subjectively, the use of the SBA algorithm was found to cause a slightly higher level of high frequency distortion in the recovered speech. However, this subjective difference was not significant at all because the amount of high frequency distortion was low and did not prove to be perceptually annoying.

The use of the SBA algorithm with the IBAP pattern of (5 4 2 2 2 0 0) implies that there are theoretically 7!/3! (7! means the factorial of 7) bit allocation patterns compared to just the 7 patterns of the VBA algorithm. However, the SNR performance of the 7/CSBC coder employing the SBA algorithm is only 0.8 dB higher than that of the 7/CSBC coder employing the VBA algorithm. This shows that as long as there are sufficient bit allocation patterns to cater for voiced, fricative and plosive sounds in speech, the use of complicated bit allocation algorithm is not really vital.

Subjective comparison between the 14/SBC/SBA coder and the 7/CSBC/VBA coders again shows that the 14/SBC coder is superior due to its finer frequency resolution. No perceptual difference could be found between the quality of the recovered speech of the 7/SBC coders and that of the 7/CSBC/VBA coders.
Table 5.5.5 Average SNRSEG measurements (in dB) for the two 7/CSBC coders

<table>
<thead>
<tr>
<th>Sentence</th>
<th>7/SCBS/VBA/AQF</th>
<th>7/CSBC/VBA/CAQF/$\phi$SBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>17.02</td>
<td>18.69</td>
</tr>
<tr>
<td>Female</td>
<td>18.63</td>
<td>20.00</td>
</tr>
<tr>
<td>Average</td>
<td>17.82</td>
<td>19.35</td>
</tr>
</tbody>
</table>

Figure 5.5.6 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the 7/CSBC/VBA/AQF coder
(3) Output noise of the 7/CSBC/VBA/CAQF/$\phi$SBA coder
Figure 5.5.7  Long-term average power spectral plots of
(1) Original speech
(2) Output noise of the 7/CSBC/VBA/AQF coder
(3) Output noise of the 7/CSBC/SBA/AQF coder
5.6 NOTE ON PUBLICATION

A paper entitled, "32 kbits/sec frequency-domain coding of wideband (0 - 7 kHz) speech" was presented at the International Conference on Digital Signal Processing in September 1984 and was published in the Conference Proceedings (168). It was written in co-authorship with Dr C S Xydeas and covers the work on 14-band subband coding described in this chapter and adaptive transform coding to be described in the next chapter. Also, U.K. patent application No. 8421498 has been made, in co-authorship with Dr C S Xydeas, to claim originality of invention of the algorithm of simplified bit allocation, described in Section 5.2.6, for the use in frequency-domain waveform coding.
A comparatively detailed study of the technique of subband coding for the transmission of wideband speech at the main bit rate of 32 kbps has been presented in this chapter. The various issues, related to the design of a subband coder, like the number of subbands, the bit allocation strategy and the quantization of the subband signals, have been examined via computer simulations to a considerable degree. A preliminary study of complex subband coding using the relatively unknown design of complex quadrature mirror filter has also been carried out. The robustness of the coder under transmission error conditions was investigated also, for the case of 14-band subband coding.

The investigation of subband coding for wideband speech at 32 kbps, began with the 7-band SBC coder employing fixed bit allocation and forward adaptive quantizer. Simulation results showed that the perceptual quality of the recovered speech was unsatisfactory due primarily to the fixed bit allocation of the scheme. To improve the performance of the 7-band coder, the bit allocation for the quantization of the subband signals has to be made adaptive to follow the time-varying spectrum of the speech signals. The use of the conventional bit allocation algorithm was found to improve considerably the SNR and perceptual performance of the 7-band coder but at the cost of an increase in system complexity. To reduce the implementation complexity of the fully adaptive bit allocation algorithm while maintaining the advantage
of having the bit assignment adaptive for improved performance, a novel simplified bit allocation algorithm was proposed and applied for the 7-band subband coder. A drop of only 0.3 dB was recorded when the 7/SBC coder employed the SBA algorithm instead of the ABA algorithm. For the speech sentences not used to derive the time-invariant bit allocation pattern of the SBA scheme, the decrease in SNR performance was 0.5 dB. More importantly, subjective listening tests showed that there was almost no difference between the recovered speech of the 7/SBC coder employing either the ABA or SBA algorithm.

Another alternative to improve the performance of the 7/SBC coder is to employ adaptive pre- and post-filtering, as in the case of ADPCM coding, to achieve finer spectral control of the output noise. Though perceptual improvement was evident from the subjective listening tests conducted, it was not significant enough to justify the additional bit rate and complexity involved. Therefore, the technique of adaptive bit allocation seemed to be more cost effective.

The preliminary study of complex subband coding showed that the correlation between the amplitude samples from adjacent bands could be exploited to improve the performance of the coder. The use of the novel amplitude quantization technique described in Section 5.5 and the simplified bit allocation strategy for the quantization of the phase signal was found to improve the SNR performance of the 7/CSBC coder by 1.5 dB compared to the same coder using the conventional quantization method. However, perceptual performance of the 7/CSBC coder was only comparable to that
of the more conventional 7/SBC coders due to the fact that their frequency resolutions are the same.

The main weakness of the 7-band coders is the coarse frequency resolution of one kHz bandwidth of each subband. Noise in each of the seven subbands is spectrally flat and thus masking of noise by the spectrally non-flat speech signals is less effective. By increasing the number of bands to 14 and by employing either the ABA or SBA algorithm, subband coding was found to produce near excellent quality recovered speech. Though the SNR performance of the 14/SBC coder at 32 kbps was lower than that of the 64 kbps 2-band SBC coder by 5.8 dB, no perceptual difference could be easily perceived. Both coders are therefore suitable for high quality wideband speech coding at their respective bit rates. Another advantage of subband coding is the ease with which it can be implemented as a variable bit rate encoder by simply varying the total number of bits assigned to the subbands.
CHAPTER 6

ADAPTIVE TRANSFORM CODING (ATC) OF

SPEECH SIGNALS
6.1 INTRODUCTION

The success of subband coding shows the advantage of coding the speech signals in the frequency-domain where adaptive allocation of bits according to perceptual criteria can be easily employed. Instead of using an analysis filter bank to convert the speech signals into frequency components, the technique of adaptive transform coding (ATC) of speech employs linear unitary transformation of a block of speech samples to achieve the similar conversion function. Once the speech samples are transformed into a different domain, adaptive bit allocation algorithm can be employed either to achieve the best SNR or perceptual performance. The commonly used transforms in ATC are unitary transforms like the Discrete Fourier Transform (DFT) and orthogonal transforms like the Discrete Cosine Transform (DCT), Discrete Sine Transform (DST), Karhunen-Loeve Transform (KLT), Walsh-Hadamard Transform (WHT), Discrete Slant Transform and Haar Transform (HT). A detailed study of the performance of the above transforms for the coding of narrowband speech can be found in References 71, 72 and 86. The unified treatment of filter bank analysis and synthesis and short-term Fourier transform (STFT) can be found also in much published literature (102-104).

In adaptive transform coding of speech, a block of speech samples are first normalized and transformed into a set of transform coefficients which are optimally quantized through adaptive bit allocation. At the receiver, the inverse transformation is carried out to obtain the
corresponding block of reconstructed speech samples (see Figure 6.1.1). The kind of transform, the transform block size and the bit allocation algorithm are three important design aspects of an adaptive transform coding system. The following section outlines the basic theory and issues related to transform coding.
Figure 6.1.1 System Block Diagram of an ATC Scheme
6.2 BASIC THEORY AND ISSUES RELATED TO ATC

6.2.1 The Block Transformation

Assume a block of \( N \) speech samples to be represented by a column vector \( X \). It is linearly transformed using a unitary or orthogonal \( N \times N \) matrix \( A \) to give a transform coefficient vector \( Y \), i.e.

\[
Y = AX
\]  (6.2.1)

with \( A^{-1} = (A^*)^T \) if \( A \) is unitary

or \( A^{-1} = A^T \) if \( A \) is orthogonal

\( A^T \) means the transpose of \( A \) and \( A^* \) means the complex conjugate of \( A \).

The transform coefficients \( y_i \) are independently quantized and transmitted to the receiver. Let the quantized vector of \( Y \) be \( \hat{Y} \). At the receiver, the inverse transformation of \( \hat{Y} \) gives the reconstructed sample vector \( \hat{X} \), i.e.

\[
\hat{X} = A^{-1} \hat{Y}
\]  (6.2.3)

For a unitary or orthogonal matrix, the average mean-squared distortion \( \overline{D} \), i.e. the average quantization noise energy of the coding scheme can be shown to be given by

\[
\overline{D} = \frac{1}{N} E \left[ (X - \hat{X})^T \cdot (X - \hat{X}) \right]
\]

\[
= \frac{1}{N} E \left[ (Y - \hat{Y})^T \cdot (Y - \hat{Y}) \right]
\]  (6.2.4)

The minimization of \( \overline{D} \) requires the choice of an optimum matrix \( A \) and an optimum rule for the quantization of the transform coefficients.
If $\sigma_i^2$ is the variance of the $i$th coefficient and $R (= \frac{1}{N} \sum_{i=1}^{N} R_i)$ is the given average bit rate, the optimum bit allocation has been shown in Section (5.2.5) to be given by

$$R_i = \bar{R} + \frac{1}{2} \log_2 \left( \prod_{j=1}^{N} \frac{\sigma_j^2}{\sigma_i^2} \right)^{1/N} \text{ (bits/sample)} \quad (6.2.5)$$

With the optimum bit allocation given by Equation (6.2.5), one can easily show, using Equation (5.2.23), (5.2.24) and (5.2.25), that the average distortion (169) is

$$\bar{D} = 2^{2\theta} \cdot 2^{-2\bar{R}} \left( \prod_{j=1}^{N} \frac{\sigma_j^2}{\sigma_i^2} \right)^{1/N} \quad (6.2.6)$$

where $\theta$ is a correction value that takes into account the performance of practical quantizer. Equation (6.2.6) says that the average distortion $\bar{D}$ is determined by the geometric mean of the variances, which are in turn determined by the matrix $A$. If $R_{xx}$ and $R_{yy}$ are the covariance matrices of the speech samples and transform coefficients respectively we have,

$$\det R_{yy} \leq \prod_{j=1}^{N} \sigma_j^2 \quad \text{(det: means determinant)} \quad (6.2.7)$$

for any matrix $A$ and

$$\det R_{xx} = \det R_{yy} \quad (6.2.8)$$
for any unitary or orthogonal matrix $A$. The variances $\sigma_j^2$ are the diagonal elements of $R_{yy}$. In particular, we have

$$\det R_{xx} = \prod_{j=1}^{N} \lambda_j$$

(6.2.9)

where the values $\lambda_j$ are the eigenvalues of $R_{xx}$. Equations (6.2.6) to (6.2.9) imply that a minimum distortion is achieved if the variances are equal to the eigenvalues. For the variances of $y_i$ to be the eigenvalues of $R_{xx}$, the matrix $A$ must be formed by the eigenvectors of $R_{xx}$. Such a matrix is called the Karhunen-Loeve transform (KLT) matrix. The transform coefficients are uncorrelated also when KL transform is used.

If we assume that the quantizer parameter $\delta$ of Equation (6.2.6) does not change whether we quantize samples in the time or transform domain, the average distortion of a PCM scheme is given by

$$\bar{D}_{PCM} = \sigma^2 2^2 \delta 2^{-2R} \sigma^2$$

(6.2.10)

The coding gain of a transform coding scheme over PCM is then given by

$$G_{TC} = \frac{\bar{D}_{PCM}}{\bar{D}} = \frac{\sigma^2}{\prod_{j=1}^{N} \sigma_j^2}^{1/N}$$

(6.2.11)

$G_{TC}$ is the increase in signal-to-noise ratio (SNR) over PCM. It is also the ratio of the arithmetic and geometric mean of the variances of
the transform coefficients if $A$ is unitary or orthogonal because the signal variance $\sigma^2$ is then equal to the average of the variances of the transform coefficients, i.e.

$$\sigma^2 = \frac{1}{N} \sum_{j=1}^{N} \sigma_j^2 \quad (6.2.12)$$

Though KL transform is theoretically the optimum transform it is not a practically attractive solution because of the need to determine the time-varying speech covariance matrix $R_{xx}$ and its associated eigenvectors which have to be transmitted to the receiver also.

To overcome these problems, time-invariant but suboptimum transform like the Discrete Cosine Transform (DCT), Discrete Sine Transform (DST), Walsh-Hadamard Transform (WHT), Discrete Slant Transform and Haar Transform (HT) can be used instead. Their performances in the applications of speech and video coding to achieve data compression are well reported in the literature \cite{71,72,84,85}.

To give some insight into the results that can be obtained using the various transforms, Zelinski and Noll \cite{71} model the speech signals as a stationary tenth-order Markov process whose first ten autocorrelation coefficients are identical to that of the speech signals sampled at 8 kHz. Knowing $R_{xx}$ (obtained from the ten coefficients) and the transform matrix $A$, the covariance matrix $R_{yy} = A R_{xx} A^T$ in the transform domain can be obtained. Thus, the coding gain given by Equation (6.2.11) can be calculated. Figure 6.2.1 shows the coding gain against blocksize
N for the optimum KLT and four suboptimum transforms. Among all the suboptimum transforms, the DCT has been found to have a performance closest to that of the KLT in terms of coding gain. In fact, it can theoretically be derived from the KLT when the adjacent data element correlation tends to unity. The DCT of an N-point sequence \( x(n) \) is defined as

\[
x_c(k) = \sum_{n=0}^{N-1} x(n) c(k) \cos \left( \frac{2(n+1)km}{2N} \right)
\]  

(6.2.13)

for \( k = 0, 1, \ldots, N - 1 \)

and \( c(k) = 1, \) when \( k = 0 \)
\[
= \sqrt{2}, \quad k = 1, 2, \ldots, N - 1.
\]

The inverse DCT (IDCT) is defined as

\[
x(n) = \frac{1}{N} \sum_{k=0}^{N-1} x_c(k)c(k) \cos \left( \frac{(2n+1)k\pi}{2N} \right)
\]  

(6.2.14)

for \( n = 0, 1, 2, \ldots, N - 1. \)

Fast algorithms have been derived for implementing the DCT with great computational efficiency.

As can be seen from Figure 6.2.1, the coding gain \( G_{TC} \) of a transform increases generally with the transform blocksize. This implies that for the adaptive transform coding of speech to be efficient, the blocksize of the transform has to be large. The conventional ATC speech
coder usually employs DCT of blocksize equal to or greater than 128. However, the processing requirement of a transformation generally increases enormously with the transform blocksize. An example of this is the computational requirement of the DCT against the blocksize shown in Figure 6.2.2. The implementation cost of an ATC speech coder thus increases with the blocksize of the transform used.

6.2.2 Quantization of the Transform Coefficients

The quantization of the transform coefficients is of fundamental importance since it determines the accuracy of preservation of the short-term signal spectrum, and hence the quality of the recovered speech. Usually the transform coefficients are independently quantized, with the step-sizes of the quantizers and the bit allocation pattern determined from the energy distribution of the basis spectrum. For maximum SNR performance, the number of bits assigned for the quantization of the coefficients are determined by Equation (5.2.26), as in the case of subband coding. For optimum perceptual results, the bit assignment rule given by Equation (5.2.29) is used instead. In order to reduce the coder complexity due to the use of the fully adaptive bit allocation algorithm, the simplified bit algorithm described in Section 5.2.6 can be employed also.
Figure 6.2.1  Coding gains in SNR over PCM against transform blocksize for various unitary transforms (after Reference 71)

\[ \text{No. of addition} = \frac{3N}{2} (\log_2 N - 1) + 2 \]

\[ \text{No. of multiplication} = N \log_2 N - \frac{3N}{2} + N \]  (Reference 172)

Figure 6.2.2  Number of additions and multiplications for the fast DCT against transform blocksize N (after Reference 172)
6.2.3 Determination of the Basic Spectrum

As the determination of the bit allocation pattern for the quantization of the transform coefficients, critical to the performance of an ATC coder, is based on the energy distribution of the basis spectrum in the transform domain, the determination of the basis spectrum is therefore an important design issue of an ATC coder. Zelinski and Noll proposed the following procedure for the estimation of the basis spectrum (71).

The input speech vector $X$ is transformed into $Y$ and the resulting amplitudes of the transform coefficients $y_1$ are squared; the values of $y_1^2$ can differ significantly from their expected values. By averaging over $M$ neighbouring values, we obtain $L (= N/M)$ samples, $\sigma_i^2$, of the basis spectrum to be estimated (see Figure 6.2.3). The quantized versions $[\sigma_i^2]_q$ of these samples form the side-information that has to be transmitted to the decoder together with the quantized transform coefficients. The encoder and decoder use this side-information to construct a complete estimate of the basis spectrum which in turn can be used to determine the bit allocation pattern.

The complete basis spectrum is obtained from the logarithmic magnitudes of the values of the $L$ samples $[\sigma_i^2]_q$ by employing a simple straight-line interpolation as shown in Figure 6.2.3d. The interpolation of the logarithmized values of the samples ensures a smoother approximation of the basis spectrum. It should be noted that the estimation procedure
outlined above is based on the assumption that the basis spectrum has a smooth shape which is not always the case. Beside being used for the determination of the bit allocation algorithm, the \( N \) estimated variances of the basis spectrum are used also to adjust the amplitude ranges of the \( N \) quantizers of the ATC coder.
Figure 6.2.3 Estimation of the basis spectrum (example: $N = 12$, $L = 4$, $M = 3$).

(a) Basis spectrum of speech
(b) Actual squared magnitudes of the transform coefficients
(c) Averaged samples derived from (b)
(d) Estimated basis spectrum obtained by interpolation using the quantized version of the averaged samples of (c)
6.3 THE PINNED-KL AND PINNED SINE TRANSFORM

For the adaptive transform coding of video signals, transform blocksize is normally limited to 16 or less. However, blocksize of 128 or more is usually used for adaptive transform coding of speech signals. This is because coding gain decreases with the decrease in transform blocksize but asymptotically reaches the KLT limit when the blocksize is more than 128 for speech signals. Large blocksize ATC of speech is a complex system and generally requires an array processor for real-time implementation. If the quality requirement of a speech coding scheme can be relaxed, ATC using small primary blocksize might be a good candidate due to its reduced complexity. Very few research reports on the use of small blocksize ATC for speech coding can be found in the existing literature. One of the schemes proposed by Fjallbrant makes use of modified DFT of blocksize 9 with adaptive skipping and recursive quantization of transform coefficients for narrowband speech coding at 16 kbps and below. Though small blocksize ATC is rarely used for speech coding, it is used widely in the area of video coding. One of the efficient small blocksize ATC schemes used for coding video signals is the pinned KL transform (PKLT) proposed by Jain. The simplified form of PKLT is termed pinned sine transform (PST) by its originator. They are both recursive block coding techniques that exploit the interblock redundancy. PKLT was experimentally proved to achieve better SNR performance than the KLT or DCT of the same blocksize for video coding.
Since the principles of ATC of speech and video signals are similar, it would be of interest to investigate the performance of PKLT and PST for the coding of wideband speech at 32 kbps. The recovered speech quality of the ATC scheme employing small blocksize PKLT or PST is not expected to exceed that provided by large blocksize (≥ 128) ATC scheme using the DCT. However, its reduced complexity and achieved quality might justify its application. In the following sections, the form of signal representation which leads to the derivation of PKLT, the theory of PKLT and PST and the basic concept of recursive block coding technique will be represented. A speech coding scheme employing the PKLT/PST at 32 kbps for wideband speech will be described also and its performance presented.
6.3.1 Signal Representation

Let speech signal be represented by a finite one-dimensional random process \( \{x_i\} \) with zero mean and unit variance and an autocorrelation function given by

\[
E\left[x_i x_{i+n}\right] = \rho^n
\]

(6.3.1)

for \( i = 0, 1, 2, \ldots, N, N+1 \).

It is well known that the sequence \( \{x_i\} \) can be represented by a first-order stationary Markov process as

\[
x_{i+1} = \rho x_i + \epsilon_i, \quad i \geq 0
\]

(6.3.2)

with

\[
E[\epsilon_i] = 0, \quad E[\epsilon_i \epsilon_j] = (1 - \rho^2) \delta_{ij}
\]

(6.3.3)

and a suitably chosen initial condition \( x_0 \).

The above Markov process, for a fixed \( N \), can be represented also by the following equations:

\[
x_i = \alpha (x_{i+1} + x_{i-1}) + \nu_i, \quad 1 \leq i \leq N
\]

(6.3.4)

\[
x_0 = \rho x_1 + \nu_0
\]

(6.3.5)

\[
x_{N+1} = \rho x_N + \nu_{N+1}
\]

(6.3.6)

where \( \alpha = \rho/(1 + \rho^2) \) and \( \{\nu_i, i = 0, 1, \ldots, N, N+1\} \) is a well-defined
351

random process. Its autocorrelation function can be shown (Appendix V) to be given by

\[ E[v_i v_j] = \begin{cases} 
\beta_2^2 \frac{1 - \rho^2}{1 + \rho^2}, & i = j, \\
- \alpha \beta_2^2, & |i-j| = 1 \\
0, & \text{otherwise,}
\end{cases} \]  \hspace{1cm} (6.3.7)

for \( i, j = 1, \ldots, N \) and at the endpoints

\[ E[v_0 v_i] = \beta_2^2 \frac{1 - \rho^2}{1 + \rho^2} = E[v_{N+1}^2], \]  \hspace{1cm} (6.3.8)

\[ E[v_0 v_1] = - \alpha \beta_2^2 = E[v_N v_{N+1}], \]  \hspace{1cm} (6.3.9)

\[ E[v_i v_j] = E[v_{N+1} v_j] = 0; \ i > 1, j < N. \]  \hspace{1cm} (6.3.10)

### 6.3.2 The Pinned-KL Transform (PKLT)

If \( X \) is a one-dimensional \( N \times 1 \) vector with autocorrelation matrix \( R_{xx} \), the KL transform of \( X \) is a matrix \( \Phi \), composed of the eigenvectors of \( R_{xx} \) and is defined by the relation

\[ \Phi^T R_{xx} \Phi = \Gamma \]  \hspace{1cm} (6.3.11)

where \( \Gamma \) is a diagonal matrix of eigenvalues \( \sigma_i^2 \).

From equation (6.3.7) to (6.3.10), if we define \( \Psi^* \) as an \((N+2) \times 1\) vector with elements \( \{v_0, v_1, \ldots, v_N, v_{N+1}\} \), its \((N+2) \times (N+2)\) autocorrelation matrix \( C^* \) can be written as
Next consider an N×1 vector \( V \) with elements \( \{v_1, v_2, \ldots, v_N\} \). The N×N autocorrelation matrix \( C \) of \( V \) is given by

\[
C = E[VV^T] = \beta_2^2 \begin{pmatrix}
1 & -\alpha & 0 & \ldots & 0 \\
-\alpha & 1 & -\alpha & \ddots & \\
0 & -\alpha & 1 & \ddots & \\
& & & \ddots & -\alpha & 1 \\
0 & \ldots & 0 & -\alpha & 1
\end{pmatrix}
\]

(6.3.13)

\[
\Delta = \beta_2^2 Q
\]

(6.3.14)

The matrix \( Q \) is a symmetric tridiagonal Toeplitz matrix.

The eigenvectors \( \psi_{ij} \) and the eigenvalues \( \sigma_i^2 \) of the N×N symmetric tridiagonal Toeplitz matrix \( Q \) are given by

\[
(177)
\]
\[ W_{ij} = \sqrt{\frac{2}{N+1}} \sin \left( \frac{i\pi}{N+1} \right) \] (6.3.15)

and

\[ \sigma_i^2 = 1 - 2\alpha \cos \left( \frac{i\pi}{N+1} \right) \text{ for } i = 1, \ldots, N. \] (6.3.16)

Since \( \bar{\theta}^2 \) is a scaling constant, the eigenvectors of \( C \) are given also by the Equation (6.3.15).

Next, consider a first-order stationary Gauss-Markov sequence \( \{x_i, i = 0, 1, \ldots, N, N+1\} \). Let the two given boundary elements be

\[ x_0 = c, \quad x_{N+1} = d \] (6.3.17)

If \( X \) and \( V \) are defined as \( N \times 1 \) vectors of elements \( \{x_1, \ldots, x_N\} \) and \( \{v_1, \ldots, v_N\} \) respectively, Equation (6.3.4) can be written as

\[ QX = V + B \] (6.3.18)

where \( Q \) is the \( N \times N \) tridiagonal matrix in Equation (6.3.14) and \( B \) is an \( N \times 1 \) vector containing only the information at the end points, i.e.,

\[ b_1 = ac, \quad b_N = ad, \]

\[ b_k = 0 \quad \text{for } 2 \leq k \leq N-1 \] (6.3.19)

Since \( c \) and \( d \) are given and \( V \) and \( B \) are uncorrelated (Appendix V), the expected value, \( U \), of \( X \) given \( B \) is
\[ U \triangleq E \left[ X/B \right] \]
\[ = \begin{bmatrix} Q^{-1} \end{bmatrix} \begin{bmatrix} B + E \left[ Y/B \right] \end{bmatrix} \]
\[ = \begin{bmatrix} Q^{-1} \end{bmatrix} B \]
\[ \triangleq X_B \] (6.3.20)

The autocorrelation matrix of the difference vector \( Y = X - U \) is then given by

\[ R_{yy} = E \left[ (X - U)(X - U)^T / B \right] \]
\[ = \begin{bmatrix} Q^{-1} \end{bmatrix} E \left[ vv^T \right] Q^{-1} \]
\[ = \beta_2^2 Q^{-1} \] (6.3.21)

Hence, the KL transform matrix of the difference vector \( Y \) is composed also of the eigenvectors defined by Equation (6.3.15). It shows that the KL transform matrix of the difference vector obtained in the way described above is time-invariant. The time-varying autocorrelation factor \( \rho \) associated with the conventional KL transform is still required to be known both at the transmitter and receiver to form \( Q^{-1} \) for the estimation of \( X \) given the two endpoints \( c \) and \( d \). KL transform applied in this specific procedure is known as the pinned KL transform.
6.3.3 Recursive Block Coding Technique

In the conventional transform coding methods, successive blocks of input data are assumed to be independent. This is fairly accurate if the block size is large and interblock correlation can be ignored. However, if the block size is small, it may be desirable to exploit the interblock redundancy. The noncausal decomposition of $X$ into $Y$ and $X_B$ by

$$Y = X - X_B$$  \hspace{1cm} (6.3.22)

where $X_B$ is defined by (6.3.20) can be employed to design a recursive block coding scheme.

Figure 6.3.1 shows the realization of the noncausal decomposition of $X$ using Equation (6.3.22). A recursive block coder based on the noncausal decomposition is shown in Figure 6.3.2. In coding the $(K+1)$th block of $N+1$ samples, the first sample $x_0$ comes from the $(N+1)$th sample of the $K$th block which has already been coded. Hence for each successive block, one only need to quantize the $x_{N+1}$ sample of that block and the transform coefficients of the difference vector $Y$. The interblock redundancy is exploited in a recursive block coder because the last elements of the present and the previous block, i.e. $x_{N+1}$ and $x_0$, are used to form an estimate of $X$. When the block size is large, this configuration of recursive block coding may not be efficient any more.
Figure 6.3.1  Noncasual decomposition realization

Figure 6.3.2  System block diagram of a recursive block coder (RBC)
6.3.4 The Pinned-Sine Transform (PST)

A recursive block coder using the pinned-KL transform of blocksize 8 was found to outperform the ATC coder employing either the conventional KLT or DCT of the same blocksize for video signals. Complexity of PKLT arises from the need to determine the value of \( p \) and the matrix \( Q^{-1} \). As the role of \( Q^{-1} \) is to estimate \( X \) given the two endpoints, a reasonable alternative to approximate the function of \( Q^{-1} \) is linear interpolation. Assume that \( p \) has a value of 0.9, the vector \( X_B \) in Equation (6.3.20) is numerically given by

\[
X_B = Q^{-1} B
\]

\[
= \begin{bmatrix}
0.863c + 0.096d \\
0.735c + 0.194d \\
0.615c + 0.293d \\
0.503c + 0.396d \\
0.396c + 0.503d \\
0.293c + 0.615d \\
0.194c + 0.735d \\
0.096c + 0.863c
\end{bmatrix}
\]  \hspace{1cm} (6.3.23)

If linear interpolation is used to estimate \( X \) given \( c \) and \( d \), then \( X_B \) is given by
The difference vector obtained via the simplified method of linear interpolation no longer has the properties that its KLT is simply the sine transform given by Equation (6.3.15). If the PKLT recursive block coder configuration is retained with the replacement of the proper noncausal decomposition procedure by linear interpolation, a simplified system is obtained and it is termed the PST recursive block coder. If the correlation of the input samples is high, the reduction in transform efficiency of the simplified scheme is not expected to be high.
In the investigation of adaptive transform coding of wideband speech at 32 kbps, an ATC scheme employing a blocksize 128 DCT was simulated. The system design is similar to that proposed by Zelinski and Noll described in Section 6.2 except for the following modifications. The 16 primary coefficients (i.e. \( L = 16 \)), used for the estimation of the basis spectrum, are calculated every 8 msec and the average of two sets of these coefficients, from adjacent frames, is used to define the step-sizes of the transform coefficient quantizers and the bit allocation pattern. A 3-bit Gaussian quantizer is used to quantize the 16 primary values of the average basis spectrum and the quantized values are transmitted to the receiver. Normalization of the input samples is also carried out every 256 samples. The normalization parameter is quantized using a 5-bit Gaussian quantizer.

For the adaptive bit allocation for the quantization of the transform coefficients, both the fully adaptive bit allocation algorithm given by Equation (5.2.29) with \( \gamma = -0.2 \) and the simplified bit allocation algorithm described in Section 5.2.6 are tested. The IBAP pattern of the SBA algorithm is determined as:

\[
\{1 \times 7, 4 \times 6, 5 \times 5, 9 \times 4, 20 \times 3, 25 \times 2, 28 \times 1, 36 \times 0\},
\]

that is, out of the 128 transform coefficients, 1 coefficient is quantized with 7 bits, 4 coefficients with 6 bits etc. Once the bit allocation pattern and the quantizers' step-sizes are determined, the transform coefficients are quantized by Laplacian AQFs and multiplexed.
together with the side-information, which consists of the normalization parameter and the step-sizes, for transmission to the receiver.

6.4.1 Smoothing Techniques to Reduce the Effect of Interblock Discontinuities

In general, some degradation in the quality of the recovered speech of an ATC coder at medium and low bit rate is caused by inter-block discontinuities due to the block processing nature of such a coder. Though the underlying speech can be very good, the 'clicking' noise arising from interblock discontinuities is perceptually unacceptable. This form of distortion is perceptually quite different from the 'hissing' and 'rumbling' noise existing in ADPCM coders. One suggested solution to this problem is to apply 10% overlap between adjacent blocks as shown in Figure 6.4.1(67). Another method is to employ either a median filtering or moving average filtering process to a few samples at both ends of each block(179). All three methods were investigated. The 10% overlap scheme was found to be the least effective in reducing interblock discontinuities because fewer bits are available for the quantization of the transform coefficients which in turn increases the amount of block-end distortion. The method of median filtering was found to give some subjective improvement while the best performance was obtained from the moving average method. In its use, 10 samples, \( x_1, x_2, \ldots, x_{10} \) (last 5 samples of the previous block and first 5 samples of the present block) were replaced by \( y_1, y_2, \ldots, y_{10} \) where
\[ y_i = \frac{1}{3} (x_{i-1} + x_i + x_{i+1}) \], and \( i = 1, \ldots, 10 \). Figures 6.4.2a and 6.4.2b show the frequency spectra of the output noise with and without smoothing respectively. This technique substantially improves the subjective quality of the recovered speech.

Figure 6.4.1 Inter-block overlapping with trapezoidal windowing for DCT analysis/synthesis
Figure 6.4.2a  Frequency spectrum of 80 msec of output noise of the ATC coder without smoothing operation

Figure 6.4.2b  Frequency spectrum of 80 msec of output noise of the ATC coder with smoothing operation
6.4.2 Results and Discussions

The average SNRSEG performance of the large blocksize ATC wideband speech coder employing the fully adaptive or simplified bit allocation algorithm is given in Table 6.4.1. Compared to the 14/SBC/SBA coder described in the previous chapter, the ATC/ABA coder is 6 dB lower in terms of SNRSEG measurement. A similar SNR difference of about 6 dB between SBC and ATC is reported also in Reference 82 for the coding of narrowband speech at 2 bits/Nyquist sample. However, it should be noted that the perceptual difference between the quality of the recovered speech of ATC and that of SBC is not as significant as suggested by the large difference in SNR values. Informal subjective listening tests show that the 14/SBC is better than the ATC/ABA coder. The recovered speech of the latter is still degraded by a slight distortion due to interblock discontinuities despite the application of the smoothing technique to the block-end samples. The output noise spectra of both coders are shown in Figure 6.4.3. The fine spectral control of the output noise of the ATC coder can be clearly seen from the spectral plot.

The ATC coder employing the SBA algorithm is 1 dB below in SNR compared to the same coder employing the ABA algorithm. The block-end distortion of the former is more pronounced and the recovered speech is degraded also by a "whispeary" noise. This means that as the noise level in ATC/ABA, at this bit rate, is just at the threshold of audibility, the
use of the full ABA algorithm becomes necessary. However, if more bits are allowed for the ATC coder, the SBA algorithm might prove to be a valuable method for reducing the coder complexity.
Table 6.4.1 Average SNRSEG measurements (in dB) for the large blocksize ATC using the ABA and SBA algorithm

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Coder</th>
<th>ATC/ABA</th>
<th>ATC/SBA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>14.60</td>
<td>13.73</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>12.58</td>
<td>11.50</td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td>13.59</td>
<td>12.62</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.4.3 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the ACT/ABA coder
(3) Output noise of the 14/SBC/AQF/SBA coder
6.5 SMALL BLOCKSIZE ATC WIDEBAND SPEECH CODERS EMPLOYING THE PINNED-KL OR PINNED SINE TRANSFORM

Prompted by the fact that an ATC scheme employing small transform blocksize is generally less complex and more amenable to real-time implementation than the same scheme employing large transform blocksize, an ATC wideband speech coder employing blocksize 9 pinned-KL transform or pinned-sine transform was investigated. Figure 6.5.1 shows the system block diagram of an ATC speech coder employing either the PKLT or PST. The input speech is first normalized every 216 (= 3 x 72) samples and voiced/unvoiced segments of speech are treated using two slightly different procedures. The voiced/unvoiced decision, based on energy and zero crossing thresholds, is made every 216 samples. When both the conditions that the energy is greater than a certain threshold $E_T$ and the number of zero crossings within a 216-sample block is below a certain threshold $Z_T$, voiced speech is deemed present. Otherwise, that block of samples is treated as unvoiced speech. The energy of a block of samples can be easily derived from the parameter that is used to normalize the input samples. Thus no additional complexity is incurred. Also, the number of zero crossings can be obtained by just counting the sign changes of the samples.

Voiced speech

The normalized input speech signal is divided into successive blocks of 9 samples. Consider the operation of the coder during the $i$th block of normalized input samples. The last elements of the previous
Figure 6.5.1  System block diagram of a Pinned Sine Transform speech coder
and present blocks, i.e. \( x_{(i-1)9} \) and \( x_{19} \), are quantized independently using a 6 bit Laplacian quantizer. The quantizer samples \( \hat{x}_{(i-1)9} \) and \( \hat{x}_{19} \) are used to form an estimate \( \hat{X}_i = \{ \hat{x}_{i1}, \ldots, \hat{x}_{i8} \} \) of the actual normalized input vector \( X_i = \{ x_{i1}, \ldots, x_{i8} \} \) using \( \alpha Q^{-1} \) described in Section (6.3) for the case of PKLT or linear interpolation for the case of PST. \( \hat{X}_i \) is subtracted from \( X_i \) to form the residue vector \( Y_i = \{ y_{i1}, \ldots, y_{i8} \} \). \( Y_i \) is then sine transformed to yield the transform coefficient vector \( U_i = \{ u_{i1}, \ldots, u_{i8} \} \) which is then quantized to \( \hat{U}_i = \{ \hat{u}_{i1}, \ldots, \hat{u}_{i8} \} \) using Gaussian AQFs.

The bit allocation pattern and the transform coefficient quantizer step-sizes are derived once every eight transform vectors \( U_i \). The variances \( \sigma^2_1, \sigma^2_3, \sigma^2_5 \) and \( \sigma^2_7 \) of the \( u_{i1}, u_{i3}, u_{i5} \) and \( u_{i7} \) coefficients are defined as

\[
\sigma^2_j = \frac{1}{8} \sum_{k=i}^{i+7} u_{kj}^2
\]

where \( u_{ij} \) is the \( j \)th coefficient of the \( i \)th block. It should be noted that this procedure of estimating the actual variances of the transform coefficients by calculating from 8 successive vectors \( U_i \) is different from the basis spectrum estimation used in the conventional ATC scheme. The \( \log_2 \) of these variances are then quantized using 4-bit Gaussian quantizer. The variances \( \sigma^2_2, \sigma^2_4, \sigma^2_6 \) and \( \sigma^2_8 \) of the coefficients \( u_{i2}, u_{i4}, u_{i6} \) and \( u_{i8} \) are estimated by linearly interpolating the quantized \( \log_2 \) variances. The quantized and estimated variances are...
then used to determine the bit allocation pattern using the full ABA algorithm given by Equation (5.2.29) with $\gamma = -0.7$. They are used also to define the transform coefficient quantizer step-sizes.

**Unvoiced speech**

When unvoiced speech is deemed present in the input signal, the linear interpolation step employed to obtain the residue vector is omitted. The block-end elements are quantized using a 2-bit Gaussian quantizer. The normalized input vectors are sine transformed and quantized using Gaussian AQF. The bit allocation pattern and the transform coefficient quantizer step-sizes are determined in the same way as that described for voiced speech.

The bit allocations of the system for a block of 72 (= 8 x 9) samples for voiced and unvoiced speech are summarized in Table 6.5.1 and 6.5.2 respectively. Gaussian AQF is employed for the quantization of the transform coefficients because the most important transform coefficient $u_{11}$ has an amplitude histogram close to the Gaussian pdf (see Figure 6.5.2).

The final system bit rate is slightly higher than 32 kbps because of the need to transmit the normalization constant and the voiced/unvoiced decision. This incurs an additional bit rate of 445 bit/sec if 5 bits are assumed for the quantization of the normalization constant and 1 bit is required for the voiced/unvoiced decision. The total bit rate of the system is now 32.45 kbps.
<table>
<thead>
<tr>
<th>End elements ((x_{29}))</th>
<th>Variances (\sigma_1^2, \sigma_3^2, \sigma_5^2, \sigma_7^2)</th>
<th>8 vectors of (U_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of samples</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Bit allocation</td>
<td>6/element</td>
<td>4/element</td>
</tr>
<tr>
<td>No. of bits</td>
<td>48</td>
<td>16</td>
</tr>
</tbody>
</table>

Total no. of bits = 144 = 32 kbps

Table 6.5.1 Summary of the bit allocation of the PKLT/PST coder for voiced speech

<table>
<thead>
<tr>
<th>End elements ((x_{29}))</th>
<th>Variances (\sigma_1^2, \sigma_3^2, \sigma_5^2, \sigma_7^2)</th>
<th>8 vectors of (U_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of samples</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Bit allocation</td>
<td>2/element</td>
<td>4/element</td>
</tr>
<tr>
<td>No. of bits</td>
<td>16</td>
<td>16</td>
</tr>
</tbody>
</table>

Total no. of bits = 144 = 32 kbps

Table 6.5.2 Summary of the bit allocation of the PKLT/PST coder for unvoiced speech
Figure 6.5.2 Amplitude histograms of the transform coefficients $u_{i1}$, $u_{i2}$ and $u_{i3}$
6.5.1 Results and Discussions

Non real-time computer simulations of ATC wideband speech coders using the pinned-KL transform and pinned-sine transform were carried out using the two speech sentences listed in Section 3.7 as input data. The average SNRSEG measurements of both systems are given in Table 6.5.3.

It can be seen from the SNR measurements that the use of the simplified PST only results in a drop of 0.7 dB compared to the same coder using the more complicated PKLT. Informal subjective listening tests did not reveal any significant difference between the quality of the recovered speech of the PST and PKLT coder. This suggests that the complicated estimation procedure of $aQ^{-1}$ used in PKLT can be replaced by the straight-line interpolation procedure used in PST.

The long-term average spectral plot of the output noise of the PST coder is shown in Figure (6.5.3) in comparison with that of the original speech spectrum and the output noise of the large blocksize ATC/ABA coder at the same bit rate. As can be seen from Figure 6.5.3, the PST coder only achieves a coarse shaping of the quantization noise in contrast to the finer noise control of the ATC/ABA coder. Though the SNR performance of the PST coder is higher than that of the ATC/ABA coder by about 4.5 dB (see Table 6.4.1), the perceived level of noise in its recovered speech is higher than that in the latter. This is because fine shaping of noise in large blocksize ATC leads to more effective masking of noise by the speech spectrum. The coarse control
Table 6.5.3 The average SNRSEG performance (in dB) of the ATC coder using the PKLT or PST

<table>
<thead>
<tr>
<th>Sentence</th>
<th>PKLT</th>
<th>PST</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>17.51</td>
<td>17.00</td>
</tr>
<tr>
<td>Female</td>
<td>20.15</td>
<td>19.22</td>
</tr>
<tr>
<td>Average</td>
<td>18.83</td>
<td>18.11</td>
</tr>
</tbody>
</table>

Figure 6.5.3 Long-term average power spectral density plots of
(1) Original speech
(2) Output noise of the ATC/ABA coder
(3) Output noise of the PST/ABA coder
of noise in the PST coder is due to the small blocksize nature of the coder. However, unlike the case of large blocksize ATC, the small blocksize PST coder with the independent quantization of the pinned-elements does not suffer from blockend distortions due to interblock discontinuities. Therefore, no selective smoothing of blockend elements is required for the PST coder. The main perceptual distortion was found to be in the form of a low level audible 'hissing' noise. Further comparisons between the PST coder and other time and frequency-domain coders will be given in the following sections.
6.5.2 Comparison Between the 32 kbps Small Blocksize PST and Large ATC Coders with the 64 kbps 2-band SBC Coder

The average SNRSEG performance of the 32 kbps PST wideband speech coder is about 6.9 dB lower than that of the 64 kbps 2-band SBC coder. The long-term average spectral plots of the output noise of the two coders (see Figure 6.5.4) also show a gap of about 7 dB between the two curves throughout the whole frequency range. However, the perceptual difference between the recovered speech of the two coders is not as great as that suggested by their big SNR difference. The 6.9 dB reduction is in fact much lower than the expected 12 dB reduction if the same 2-band SBC coder is to operate at 32 kbps. Though not as successful as the 14/SBC/SBA coder, which offers as good quality at 32 kbps as that of the 2-band SBC coder at 64 kbps, the reduced complexity of the PST coder and the reasonable quality it achieves might render it a possible candidate for the coding of wideband speech at 32 kbps.

The average SNRSEG difference between the 32 kbps ATC/ABA and the 64 kbps 2-band SBC coders has an even greater value of 11.4 dB. Nevertheless, informal subjective listening tests show that the perceptual quality of the recovered speech of the ATC scheme is only slightly worse than that of the 2-band SBC scheme. The very slight 'clicking' noise in the recovered speech of the ATC coder due to inter-block discontinuities proves to be unacceptable for high quality wideband application. It can be seen also in Figure 6.5.4 that the coarse noise shaping in
Figure 6.5.4  Long-term average power spectral density plots of

(1) Original speech
(2) Output noise of the ATC/ABA coder
(3) Output noise of the PST/ABA coder \{ at 32 kbps
(4) Output noise of the 2-band SBC coder at 64 kbps
the case of 2-band SBC coder requires the overall noise level to be much lower than the speech spectrum for the auditory masking of the noise by speech to be effective. Though the noise level of the ATC coder is relatively higher than that of the 2-band SBC coder, the finer spectral shaping of the ATC output noise makes the job of auditory masking easier. This shows the importance of noise spectral shaping to improve the perceptual quality of the recovered speech at medium bit rate. It shows also that SNR measurement can only be used as a complementary performance 'yardstick'.
6.6 COMPARISONS OF THE 32 KBPS TIME AND FREQUENCY-DOMAIN WAVEFORM CODERS AND THE 64 KBPS 2-BAND SBC CODER

The various 64 and 32 kbps wideband speech coding schemes employing the various time and frequency-domain waveform coding techniques have been investigated in the past and present chapters. This section summarises and discusses the SNR and subjective performances of seven coding schemes selected from all the schemes investigated so far. For the category of 64 kbps coders, the 2-band SBC scheme employing ADPCM encoders with pre- and de-emphasis for the lower band is chosen for comparison because it offers the best performance among all the 64 kbps coders. In the category of time-domain waveform coding at 32 kbps, an ADPCM coder employing a 6th-order adaptive lattice predictor and the one-word memory Jayant's quantizer was simulated, in addition to the various ADPCM/ANS coders investigated, for comparison. The ADPCM coder also employs pre- and de-emphasis as a form of noise spectral shaping procedure. The leakage constants of the AQJ and the adaptive lattice predictor were set to $1 - 2^{-6}$ and $1 - 2^{-4}$ respectively. The adaptation constant of the latter was set to $2^{-6}$.

In the category of frequency-domain waveform coding, two 14-band SBC/AQF coders, one employs the fully adaptive bit allocation algorithm and the other the simplified bit allocation algorithm, are chosen for comparison as they offer the best subband coding performance. The other equally good subband coder with adaptive pre- and post-filtering is not considered in this comparison because it has a much higher bit
rate of 35.7 kbps. Also chosen for comparison are the large blocksize ATC coders employing the ABA and SBA algorithms and the small blocksize PST coder employing the ABA algorithm.

The average SNRSEG performance of all the seven coders are summarised in Table 6.6.1. The 64 kbps scheme of course has the higher SNR performance (24.99 dB) because it is operating at the bit rate double that of the rest. The SNR value of the 14/SBC/ABA has the next highest value of 19.88 dB, only about 5 dB below that of the 64 kbps coder despite the fact that the average number of bits available for quantization per sample is 2 bits less. The 14/SBC/SBA has the third highest SNR measurement, followed by the PST, ADPCM/PDE, ATC/ABA and ATC/SBA.

It should be noted that the SNR performance of all the seven coders can be increased beyond the measured values when the factor of noise spectral shaping is not included into their designs. However, it is well known that the SNR performance should not be the only criterion in the design of a speech coding scheme. The comparison of the five out of the seven coders in terms of the long-term average spectral density plots of their output noise for one sentence of male speech are given in Figure 6.6.1. The 64 kbps 2-band SBC coder obviously has the lowest level of noise followed by that of the 14/SBC/SBA scheme. The ATC/ABA can be seen to have the finest spectral control of the noise and thus its recovered speech has a quality only next to the 2-band SBC and 14/SBC coders. The ADPCM/PDE coder suffers from both a high level and coarse shaping of the quantization noise.
<table>
<thead>
<tr>
<th>Bit Rate kbps</th>
<th>Coder</th>
<th>Male Speech (average of 2 sentences) (dB)</th>
<th>Female Speech (average of 2 sentences) (dB)</th>
<th>Total Average (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>2-Band SBC</td>
<td>22.53</td>
<td>27.44</td>
<td>24.99</td>
</tr>
<tr>
<td>32</td>
<td>ADPCM/PDE</td>
<td>13.55</td>
<td>18.02</td>
<td>15.79</td>
</tr>
<tr>
<td>32</td>
<td>14/SBC/ABA</td>
<td>19.26</td>
<td>20.52</td>
<td>19.88</td>
</tr>
<tr>
<td>32</td>
<td>14/SBC/SBA</td>
<td>18.79</td>
<td>19.68</td>
<td>19.24</td>
</tr>
<tr>
<td>32</td>
<td>ATC/ABA</td>
<td>14.60</td>
<td>12.58</td>
<td>13.59</td>
</tr>
<tr>
<td>32</td>
<td>ATC/SBA</td>
<td>13.73</td>
<td>11.50</td>
<td>12.62</td>
</tr>
<tr>
<td>32</td>
<td>PST/ABA</td>
<td>17.00</td>
<td>19.22</td>
<td>18.11</td>
</tr>
</tbody>
</table>

Table 6.6.1  The average segmental SNR performance of the various time and frequency-domain waveform coders.
Figure 6.6.1 Long-term average power spectral density plots of

(1) Output speech
(2) Output noise of the ADPCM/PDE coder
(3) Output noise of the ATC/ABA coder
(4) Output noise of the PST/ABA coder
(5) Output noise of the 14/SBC/SBA coder
(6) Output noise of the 2-band SBC coder at 64 kbps

at 32 kbps
In terms of subjective performance, the 2-band SBC coder at 64 kbps and the two 14/SBC coders at 32 kbps have almost the same near-excellent quality for the recovered speech. The 2-band SBC coder achieves the desired design objective by having a relatively lower level of noise and a coarser control of the noise spectrum. Having to live with a higher level of noise in the case of 32 kbps, the 14/SBC coder achieves the desired design objective by having more bands and thus a finer and more effective control of the output noise to fully exploit the effect of auditory masking properties of the ear. The next scheme, in order of merit, is ATC employing the ABA algorithm. Though distortion due to inter-block discontinuities is substantially reduced by smoothing, a slightly audible level of 'clicking' noise due to discontinuities is still detectable. The use of the SBA algorithm in ATC increases the level of distortion due to inter-block discontinuities and the recovered speech is further degraded by a low but audible level of 'whispery' noise. The less complicated small blocksize PST coder has a slightly lower level of noise in the recovered speech than that of the large blocksize ATC scheme. The small blocksize nature of the PST coder makes the job of noise spectral control more difficult. The worst among all is the ADPCM/PDE at 32 kbps. The use of pre- and de-emphasis filters reduces the high frequency 'hissing' noise, otherwise present in the output speech signals, at the expense of an increase in low frequency 'rumbling' noise.
Complexity Considerations:

The advent of digital signal processing devices has facilitated the real-time implementation of a number of otherwise difficult to implement speech coding algorithms. A 2-band subband coder for example, can be implemented using a single DSP device. Coding algorithms with much higher system complexity than 2-band SBC coder generally require a few DSP devices or array processor for their real-time implementations. As system complexity is also a factor in selecting a coding algorithm for a specific application, it would be incomplete without a rough estimation of this aspect of the various coding algorithms examined so far. The implementation complexity of a coder depends to an extend on the number of multiplications/divisions, additions/subtractions and the size of the RAM required for storing the intermediate variables of the coding algorithm. Table 6.6.2 shows the main computation and storage requirements, including delays, of the coders examined. Estimations of these requirements, for each of the individual processing stages of the various coding algorithms, are given in Appendix VI.

Fast algorithms for the cosine and sine transforms were assumed in deriving the computation estimates. The ATC and PST coders required also additional $\log_2$ and $\log_2$ look-up tables. For the subband coders, the higher stages of the filterbank can be implemented using lower order FIR filters to reduce memory size, computation requirements
and coder delay. For the complex subband coder, the processing
requirement for the calculation of the square roots and the $\tan^{-1}$
are not included in the estimation. Also, excluded in the estimation
is the memory size required for the programme instructions of the
coding algorithms.

<table>
<thead>
<tr>
<th>Coder</th>
<th>Memory size (words)</th>
<th>No. of $x/\div$ (/sample)</th>
<th>No. of $+/-$ (/sample)</th>
<th>System Delay (msec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADPCM</td>
<td>50</td>
<td>37</td>
<td>38</td>
<td>0</td>
</tr>
<tr>
<td>7/SBC/SBA</td>
<td>430(370)</td>
<td>48(39)</td>
<td>47(38)</td>
<td>30(26)</td>
</tr>
<tr>
<td>14/SBC/SBA</td>
<td>650(500)</td>
<td>62(45)</td>
<td>61(44)</td>
<td>45(33)</td>
</tr>
<tr>
<td>7/CSBC</td>
<td>650(500)</td>
<td>63(46)</td>
<td>62(45)</td>
<td>45(33)</td>
</tr>
<tr>
<td>ATC/ABA</td>
<td>900</td>
<td>9</td>
<td>14</td>
<td>25</td>
</tr>
<tr>
<td>PST/ABA</td>
<td>300</td>
<td>6</td>
<td>7</td>
<td>19</td>
</tr>
</tbody>
</table>

Note: Figures in the ( ) are for the coders which use lower
order FIR filters for the higher stages of their QMF
filter banks.

Table 6.6.2 Estimates of the computation and storage requirements
of the various coders
6.7 SUMMARY AND CONCLUSION

With the continually dropping cost of digital electronics and the advent of digital signal processing devices, complicated speech coding techniques like vocoding and adaptive transform coding are once again receiving increasing interest. Though there are a number of publications in the existing literature on adaptive transform coding of narrowband speech signals at various transmission bit rates, the application of ATC for wideband speech at 32 kbps has never been reported according to the author's knowledge. This chapter therefore begins with the study of the conventional ATC technique using large blocksize (128) DCT for the coding of wideband speech at 32 kbps. Computer simulation results confirmed that very fine noise spectral control can be achieved using such a coder. However, at the average bit rate of 2 bits/Nyquist sample the distortion due to interblock discontinuities was found to be perceptually unacceptable. The problem of interblock discontinuities for ATC of narrowband speech at 16 kbps was reported also in two recent publications (82,180). The earlier papers by Zelinski and Holl (71) did not suggest the existence of this problem because side-information was not quantized and all the available bits were used for the quantization of the transform coefficients and therefore the coder was operating at a higher bit rate than it actually meant to be. To reduce the effect of interblock discontinuities, the use of the technique of selective smoothing via moving average filtering was found to be effective in improving the
perceptual quality of the recovered speech. When compared with the same category of frequency-domain waveform coder like a subband coder, ATC was found to be inferior in both SNR and perceptual performance. The technique of simplified bit allocation algorithm was incorporated also into the ATC scheme and its effectiveness examined. It reduces the SNR performance of ATC by 1 dB compared to the same coder employing the conventional adaptive bit allocation algorithm and subjectively, more noise was perceived also. This shows that when the level of noise power is high, the use of the ABA algorithm becomes necessary. However, the SBA algorithm might still prove to be a valuable alternative to reduce the coder complexity at higher transmission bit rate.

The second half of this chapter has provided a preliminary investigation of the use of pinned-KL and pinned-sine transform for the coding of wideband speech at 32 kbps. Computer simulation results suggest that the small blocksize adaptive transform coder using the PST is perhaps more cost effective than using the PKLT. The use of the simpler PST only reduces the SNR performance by 0.7 dB. The recovered speech of the PST coder was degraded by a slightly higher level of high frequency "hissing" noise when compared with that of the large blocksize ATC coder. However, unlike the latter, it did not suffer from the problem of interblock discontinuities. Its reduced implementation complexity compared to the conventional large block size ATC and the reasonable quality of recovered speech it achieved might justify its application for wideband speech coding.
CHAPTER 7

RECAPITULATION AND CONCLUSION
7.1 RECAPITULATION

Since the invention of the first speech digitization technique, PCM, by Reeves in 1939 and the first vocoding technique by Dudney in the same year, speech coding has never failed to arouse the interests of many research workers. This is particularly true in the last ten years as the continually dropping cost of integrated circuits and the advent of general purpose digital signal processing devices have made possible and facilitated the real-time implementations of many complicated speech coding algorithms. Traditionally, tremendous effort has been directed to the search and development of efficient techniques for toll-quality narrowband speech coding at the bit rates of 32 kbps and below. With the emergence of the ISDN and the growing demand for higher quality voice communications, part of the present attention and interests have been channelled to the search for commentary-quality wideband speech coding at 64 kbps and below. The work described in the last four chapters of this thesis is precisely a very modest attempt to compare, examine, simplify, improve and develop the various time and frequency-domain waveform coding techniques for the digitization of wideband speech at the bit rates of 64 and 32 kbps.

Chapter 3 presents a comparative study of the performances of the various simple time and frequency-domain waveform coding techniques applied to wideband speech at 64 kbps. The use of the simplest noise spectral shaping technique of pre-emphasis and de-emphasis was incorporated
into the ADPCM and the two-band SBC system to enhance the perceptual quality of the recovered speech. Subjective listening tests showed significant improvement in the quality of the recovered speech due to the use of PDE. Instead of minimizing the perceptual effect of the quantization noise through noise spectral shaping, the more conventional technique of reducing the power of the quantization noise, through the use of adaptive lattice prediction, was also considered to improve the subjective performance of the ADPCM coder. Among the four schemes examined, namely ADPCM with PDE and fixed prediction, ADPCM with adaptive lattice prediction, two-band SBC with ADPCM encoders and two-band SBC with ADPCM encoders and PDE, the ADPCM encoder employing adaptive lattice prediction and the two-band SBC coder employing ADPCM with PDE were found to produce a recovered speech almost indistinguishable from the original unprocessed speech. The robustness of both coders under the transmission error conditions was examined also and their perceptual performance was found to be 'reasonable' even at the bit error rate of $10^{-3}$. The real-time implementation of the two-band SBC scheme further demonstrated that very good quality wideband speech at 64 kbps could be easily achieved using the currently available DSP devices.

To achieve the bit rate of 56 kbps, while maintaining the quality of the recovered speech as close to that of the 64 kbps schemes as possible, the fixed prediction employed in the lower band of the 64 kbps 2-band scheme was replaced by a 6th-order adaptive lattice predictor. The number of bits allocated to the lower band was then
reduced from five to four so that 8 kbps could be spared for the transmission of data. The subjective quality of the recovered speech of the 56 kbps scheme was found to be only slightly worse than that of the 64 kbps schemes.

While the CCITT study group on wideband speech coding is currently undecided about i) the exact transmission bit rates, i.e. whether to encode wideband speech at the two bit rates of 64 and 56 kbps or at the lower bit rate of 48 kbps, and ii) the mechanism for switching from one bit rate to another, it was decided that our efforts should be channelled to the study of the various coding techniques at the even lower bit rate of 32 kbps. The aim, of course, was to maintain the subjective quality of the recovered speech while reducing the bit rate from 64 to 32 kbps. System complexity was expected to increase considerably for the lower bit rate coder.

The investigation of the 32 kbps wideband speech coding began with the study of the techniques of adaptive noise spectral shaping of the quantization noise in an ADPCM coding system. Three different noise shaping techniques, namely pre-emphasis and de-emphasis, adaptive noise spectral shaping through adaptive noise feedback filtering and forward adaptive pre- and post-filtering, were employed to enhance the perceptual quality of the ADPCM coded speech. The performances and comparisons of the various schemes are presented in Chapter 4. Through computer simulations and subjective listening tests, the ADPCM coder
employing the forward adaptive pre- and post-filtering was found to achieve the best subjective results but at the cost of a slightly higher amount of side-information. The use of adaptive feedback filtering of the quantization noise did improve the perceptual quality of the recovered speech. However, the overall level of the quantization noise of the ADPCM coder operating at 2 bits per sample was found to be too high for the coder to fully exploit the advantage of noise shaping. The recovered speech was found still to be subjectively unacceptable. The use of the simple technique of pre-emphasis and de-emphasis compared not too badly with the adaptive schemes. High frequency 'hissing' noise was successfully attenuated though at the expense of an increase in low frequency 'rumbling' noise which was partially masked by the speech signals. Though none of the ADPCM/ANS schemes succeeded in producing the desired quality at this bit rate for wideband application, this study clearly demonstrated the importance of noise spectral control in relation to the time-varying spectrum for improved subjective performance for a coding scheme.

In Chapter 5, the more promising technique of subband coding was examined in greater details. Subband coders, as all other frequency domain coders, have the advantage that the perceptually more important samples can be encoded with higher accuracy than those samples which are perceptually less significant. Also, the concept of noise spectral shaping can still be applied and there is the additional advantage that
quantization noise is confined to each band so that the masking of the speech signals in one band by the noise in another band can be prevented. The study of subband coding of wideband speech at 32 kbps was focussed on three groups of coders, namely the 7-band subband coders, the 14-band subband coders and the 7-band complex subband coders, each of which employed various proposed different bit allocation and quantization strategies. Subjective listening tests of the recovered speech of the various schemes clearly show that

1. it is necessary to apply adaptive bit allocation to achieve good quality wideband speech at 32 kbps;
2. perceptual improvement can be obtained by increasing the number of subbands from seven to fourteen;
3. the use of adaptive pre- and post-filtering can be employed, as an alternative to adaptive bit allocation, to achieve noise spectral control, and
4. in the case of complex subband coding, the correlation between the amplitude samples from adjacent bands can be exploited to achieve higher SNR performance.

In the course of study of the technique of subband coding, a novel simplified bit allocation algorithm was proposed and used as an alternative to the conventional fully adaptive bit allocation algorithm. Based on the statistics derived from the bit allocation patterns of a training sequence, the time-invariant bit allocation patterns for the
7 and 14 band subband coders at the various bit rates of 32, 24 and 16 kbps were found and used in the SBA algorithm. The drop in SNR performance due to the use of the SBA instead of the ABA algorithm was found to be minimal. Most important of all, subjective listening tests showed that the use of the SBA algorithm did not seem to cause any perceivable increase in distortion at this bit rate, suggesting that it is truly a valuable alternative to the complicated ABA algorithm.

Under ideal channel conditions the subjective quality of the recovered speech of the 14-band SBC coder employing either the ABA or SBA algorithm was found to be near excellent and indistinguishable from that of the 64 kbps 2-band SBC coder. Furthermore, the slight drop in perceptual quality of the 14-band SBC coder when switching from 32 to 24 kbps argues well that it is a good candidate for a variable bit rate speech coding system. Another 'side-product' of the investigation of the 14-band SBC coders is the 16 kbps coding of wideband speech compared to the coding of narrowband speech at the same bit rate. Though the amount of quantization and aliasing noise is higher and audible at 16 kbps for wideband speech, it nevertheless has the pleasing effect of wider bandwidth compared to the narrowband speech especially for unvoiced sounds.

For the first time, the use of the complex QMF in the design of a complex subband coder is reported in this thesis. The CQMF filter-bank
was used to divide the speech signals into uniform channels of amplitude and phase signals. The correlation between the amplitude samples from the adjacent bands was made use of for the design of a novel quantization technique. Based on the observation that during voiced speech, low/high amplitude samples in one band are very likely to correspond to low/high amplitude samples in another band, two different quantizers, instead of one in the conventional design, were designed to suit the low and high amplitude samples which had rather different amplitude distributions. Without increasing much the coder complexity due to use of the novel quantization scheme, an SNR improvement of 1.5 dB was observed.

The use of adaptive transform coding of wideband speech is reported in Chapter 6. Despite common belief, the large blocksize ATC coding of wideband speech seemed to suffer from the problem of block-end distortions due to inter-block discontinuities inherent in block processing techniques at low bit rate. The use of the technique of selective smoothing via moving average filtering reduced the subjective effect of inter-block discontinuities considerably. The recovered speech was, however, below that of the 14-band SBC coder. To avoid the implementation complexity of the large blocksize ATC coders, a small blocksize pinned-sine transform coder was considered and developed. Though the overall subjective quality of the recovered speech of the PST coder was below that of the conventional large blocksize ATC coder, the problem of inter-block discontinuities did not exist at all.
7.2 SUGGESTIONS FOR FUTURE RESEARCH

A brief survey of the literature on speech coding published in the past decade shows that considerable amount of effort has been channelled to modify the existing coding algorithms to achieve improved system performance, to simplify complicated algorithms to reduce implementation complexity or to cascade different techniques to achieve bandwidth compression. Indeed, truly significant contributions like the inventions of Linear Predictive Coding vocoder, adaptive DPCM and subband coding are hard to come by. While it may be true that further breakthrough in speech coding can be achieved only when the human hearing mechanism is fully understood as suggested by some eminent research workers, there are still questions unanswered especially when the three conflicting issues of bit rate, quality and complexity are to be considered simultaneously for the design of speech coding systems. Robustness of a coding scheme under transmission error conditions may also deserve more attention.

For the various subband coding schemes investigated in this thesis, quantization of the step-sizes of AQFs and the subband signals are performed on a sample-by-sample basis which is appropriately known as scalar quantization. One possibility to improve the quantization accuracy is to employ vector quantization of AQFs' step-sizes and the subband signals. Side-information for the transmission of step-sizes can also be reduced, if vector quantization is employed. In most subband coding schemes, adaptive bit allocation algorithm is only applied...
in the frequency domain. Once a subband is allocated a certain number of bits within a short-time interval, the bit allocation for the quantization of every sample in that band is constant throughout that time frame. This is obviously not the best possible strategy as the magnitudes of the subband signals also vary with time. If the bit allocation can also be made adaptive in the time-domain, SNR performance of the coder is very likely to be improved.

The other area that is worth further investigation is the pinned-sine transform speech coder presented in Chapter 6. In the present design, the quantization of the blockend elements employs fixed bit allocation (6 bits for the voiced segments and 2 bits for the unvoiced segments) whereas the quantization of the transform coefficients employs adaptive bit allocation. If the bit allocation for the quantization of the blockend elements can also be made adaptive, SNR improvement is envisaged. Another possibility to improve the performance of the PST coder is the use of vector quantization of the transform coefficient vectors instead of the present scalar quantization strategy. Coder complexity is expected to be increased of course.

Finally, it might be advantageous to incorporate pitch information in the design of a variable blocksize ATC scheme. If the blocksize of an ATC coder is made adaptive with respect to the time-varying pitch period of speech, the successive transform coefficient vectors should
have very high correlation which might be easily exploited to achieve high SNR performance. However, system complexity is exorbitant.

7.3 CLOSING REMARKS

The work described in this thesis represents a modest study of time and frequency-domain waveform coding of wideband speech at the bit rate of 64/32 kbps. It is hoped that the computer simulation results and comparisons of the various coders proposed can serve as a useful reference for the design of wideband speech coding systems. The proposed simplified bit allocation algorithm, though examined only for the use in subband and adaptive transform coding, might also be employed in other systems that require adaptive bit allocation strategy. While it is true that the design of the complex subband coder and the preliminary investigation of the pinned-sine transform speech coder do not show an overwhelming advantage over the conventional subband and transform coder, their modified forms, with possible improved performance, might have useful application in certain areas.

It has also been shown in this thesis that the 32 kbps fourteen-band subband coder employing the novel simplified bit allocation algorithm has almost the same subjective performance as that of the 64 kbps two-band subband coder and the recovered speech of the coder is almost indistinguishable from the original unprocessed speech. Despite common belief, ATe coding of wideband speech at 2 bits/sample suffers from blockend distortions due to inter-block discontinuities.
APPENDIX I

Derivation of the equation

\[ R(t) = \frac{2}{\pi} \sin^{-1} (r(t)) \]

where \( r(t) \) is the correlation function of two random variables \( X \) and \( Y \) with independent Gaussian distributions and \( R(t) \) is the correlation function of \( f(X) \) and \( f(Y) \) which are obtained from taking the signs of \( X \) and \( Y \) respectively.

Assuming \( X \) and \( Y \) are two independent random variables with Gaussian distribution and normalized amplitudes.

The probability that \( X \) falls in \( X, X+dX \), and \( Y \) also simultaneously falls in \( Y, Y+dY \)

\[ \frac{1}{2\pi(1-r^2)^{1/2}} e^{-\frac{(X^2+Y^2 - 2r XY)}{2(1-r^2)}} \, dx \, dy \]

where \( r(t) \) is the average value of the product \( XY \). (1)

The expression (1) is called the normal surface in statistical theory.

If a clipper is applied to \( X, Y \) to produce \( f(X), f(Y) \), the expectation of \( f(X)f(Y) \) is given by

\[ R(t) = \frac{1}{2\pi(1-r^2)^{1/2}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(X)f(Y) \, e^{-\frac{(X^2+Y^2 - 2r XY)}{2(1-r^2)}} \, dx \, dy \] (2)

In the case of extreme clipping, \( f(X) = 1 \) for \( X>0 \) and \( f(x) = -1 \) for \( X<0 \), equation (2) reduces to
\[ R(t) = \frac{1}{2\pi(1-r^2)^{\frac{1}{2}}} \left[ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\alpha} dX dY + \int_{0}^{\infty} \int_{0}^{\infty} e^{-\alpha} dX dY - \int_{0}^{\infty} \int_{0}^{\infty} e^{-\alpha} dX dY \right] \] (3)

where \( \alpha = (X^2 + Y^2 - 2rXY)/2(1 - r^2) \).

Equation (3) can be simplified by using the relation

\[ \frac{1}{2\pi(1-r^2)^{\frac{1}{2}}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\alpha} dxdy = 1 \] (4)

By introducing \( X = \rho \cos \phi \), \( Y = \rho \sin \phi \) and using equation (4), equation (3) becomes

\[ R(t) = 4 \int_{0}^{\pi/2} \int_{0}^{\infty} \frac{1}{2\pi(1-r^2)^{\frac{1}{2}}} e^{-\rho^2 (1-r \sin 2\phi)/2(1-r^2)} \rho d\rho d\phi \] (5)

Integration gives the final result

\[ k(t) = \frac{2(1-r^2)^{\frac{1}{2}}}{\pi} \int_{0}^{\pi/2} \frac{d\phi}{1-r \sin 2\phi} - 1 = \frac{2}{\pi} \sin^{-1}(r) \] (6)
APPENDIX II

Proof of Constraint on Quantization Noise Spectrum

Proof of constraint

\[ \frac{1}{f_s} \int_0^{f_s} \log \Gamma(f) df = 0 \]  
\[ (1) \]

Given

\[ \Gamma(f) = \left| \frac{1 - F(e^{j2\pi fT})}{1 - A(e^{j2\pi fT})} \right|^2 \]  
\[ (2) \]

where \( F \) and \( A \) can be expressed in the Z transform notation as

\[ 1 - F(z) = 1 - \sum_{k=1}^{N} b_k z^{-k} \]  
\[ (3) \]

\[ 1 - A(z) = 1 - \sum_{k=1}^{N} a_k z^{-k} \]  
\[ (4) \]

\( f_s \) is the sampling frequency, \( T \) is the sampling period and the roots of both \( 1 - F \) and \( 1 - P \) are assumed to be within the unit circle.

As \( 1 - F(z) \) is a polynomial of \( z^{-1} \), it can be factorised to give

\[ 1 - F(z) = 1 - \sum_{k=1}^{N} b_k z^{-k} \]

\[ = \prod_{k=1}^{N} (1 - c_k z^{-1}) \]
\[ (5) \]
or
\[
\log \left[ 1 - F(z) \right] = \frac{N}{k=1} \log(1 - C_k z^{-1})
\]  \hspace{1cm} (6)

since \(|C_k| < 1\) (all the zeros of \(1 - F(z)\) are inside the unit circle).

The function \(\log (1 - C_k z^{-1})\) can be expressed as
\[
\log (1 - C_k z^{-1}) = -C_k z^{-1} + \frac{1}{2} C_k^2 z^{-2} - \frac{1}{3} C_k^3 z^{-3} + \ldots 
\]
for \(K = 1, \ldots, N\). \hspace{1cm} (7)

Equation (6) and (7) give
\[
\log \left[ 1 - F(z) \right] = \sum_{n=1}^{\infty} d_n z^{-n} = \sum_{n=1}^{\infty} d_n e^{-j2\pi fT_n} \hspace{1cm} (8)
\]
where \(d_n = \frac{N}{k=1} C_k^n (-1)^n \left(\frac{1}{n}\right)\) \hspace{1cm} (9)

The integral of \(\log 1 - F(z)\) over the frequency range from 0 to \(f_s\) is then given by
\[
\int_0^{f_s} \log \left[ 1 - F(e^{j2\pi fT}) \right] df = \sum_{n=1}^{\infty} d_n \int_0^{f_s} e^{-j2\pi fT_n} df 
\]  \hspace{1cm} (10)

\[= 0\]

Similarly, it can be shown that
\[
\int_0^{f_s} \log \left[ 1 - A(e^{j2\pi fT}) \right] df = 0 \hspace{1cm} (11)
\]

Thus
\[
\frac{1}{f_s} \int_0^{f_s} \log \Gamma(f) df = 0.
\]
Equations relating to Complex Quadratic Mirror Filter

The modulated down-sampled signals \( v(n) \) and \( \hat{v}(n) \) of Figure (5.27) can be expressed in the z-transform notation as given in equations (5.2.19) which are repeated here

\[
V(z) = \frac{1}{4} \sum_{u=0}^{1} \left[ X((-1)^u(jz^k)) + X((-1)^u(-jz^k)) \right] H((-1)^u z^k) \tag{1}
\]

\[
\hat{V}(z) = \frac{1}{4j} \sum_{u=0}^{1} \left[ X((-1)^u(jz^k)) - X((-1)^u(-jz^k)) \right] H((-1)^u z^k) \tag{2}
\]

The two quadratic signals \( v(n) \) and \( \hat{v}(n) \) are upsampled by inserting zero between every two samples. The upsampled signals are then lowpass filtered to give \( u(n) \) and \( \hat{u}(n) \) which can be expressed in the z-transform notation as

\[
u(z) = V(z^2)H(z)
\]

\[
u(z) = \frac{1}{4} \sum_{u=0}^{1} \left[ X((-1)^u(jz)) + X((-1)^u(-jz)) \right] H((-1)^u z) H(z)
\]

\[
\hat{u}(z) = \hat{V}(z^2) H(z)
\]

\[
\hat{u}(z) = \frac{1}{4j} \sum_{u=0}^{1} \left[ X((-1)^u(jz)) - X((-1)^u(-jz)) \right] H((-1)^u z) H(z)
\]

The two signals \( u(n) \) and \( \hat{u}(n) \) are demodulated by \( \sin(2\pi f_s nT/4) \) and \( \cos(2\pi f_s nT/4) \) respectively. The demodulated signals \( w(n) \) and \( \hat{w}(n) \) can be expressed as

\[
w(z) = \frac{1}{2j} \left[ u(jz) - u(-jz) \right]
\]

\[
\hat{w}(z) = \frac{1}{2} \left[ \hat{u}(jz) + \hat{u}(-jz) \right]
\]

(3)
The reconstructed signal \( \tilde{x}(n) \) is obtained by subtracting \( \hat{w}(n) \) from \( x(n) \), i.e.

\[
\tilde{x}(z) = x(z) - \hat{w}(z)
\]  

Equations (2), (3) and (4) give

\[
\tilde{x}(z) = \frac{1}{8j} \sum_{u=0}^{1} \left\{ \left[ X((-1)^u(-z) + X(1(-1)^u(-z)) \right] H((-1)^u(jz)) H(jz) - \left[ X((-1)^u(-z)) - X((-1)^u(-z)) \right] H((-1)^u(jz)) H(kz) - \left[ X((-1)^u(-z)) - X((-1)^u(-z)) \right] H((-1)^u(-jz)) H(-jz) \right\}
\]

\[
= \frac{1}{8j} \sum_{u=0}^{1} \left[ X((-1)^u(z)) H((-1)^u(jz)) H(jz) - X((-1)^u(z)) H((-1)^u(-jz)) H(-jz) \right]
\]

\[
= \frac{1}{4j} X(z) \left[ H^2(jz) - H^2(-jz) \right]
\]  

Assume now that the lowpass filter \( H(z) \) is symmetrical FIR filter of even length \( M \). The evaluation of \( H(jz) \) and \( H(-jz) \) on the unit circle with \( j = e^{j\omega_s T/4} \) yields respectively

\[
H(j e^{j\omega T}) = H(\omega + \frac{\omega_s}{4}) e^{-j(M-1)\pi\omega/\omega_s} e^{-j(M-1)\pi/4}
\]  

\[
H(-j e^{j\omega T}) = H(\omega - \frac{\omega_s}{4}) e^{-j(M-1)\pi\omega/\omega_s} e^{j(M-1)\pi/4}
\]

where \( H(\omega) \) is the magnitude of the Fourier transform \( H(e^{j\omega T}) \). Substituting (6) and (7) into (5) gives
\[ X(e^{j\omega T}) = \frac{X(e^{j\omega T})}{4j} e^{-j(M-1)\omega T} \left[ H^2(\omega + \frac{\omega_s}{4}) e^{-j(M-1)\pi/2} ight. \\
\left. - H^2(\omega - \frac{\omega_s}{4}) e^{j(M-1)\pi/2} \right] \quad (8) \]

For \( M = 4k \), Equation (8) becomes

\[ X(e^{j\omega T}) = \frac{X(e^{j\omega T})}{4} e^{-j(M-1)\omega T} \left[ H^2(\omega + \frac{\omega_s}{4}) + H^2(\omega - \frac{\omega_s}{4}) \right] \quad (9) \]
(a) Input Decision Thresholds and Output Levels of a Gaussian Quantizer

\[ x_i = \text{Input Threshold} \]
\[ y_i = \text{Output Level} \]
<table>
<thead>
<tr>
<th>Bits</th>
<th>( x_i )</th>
<th>( y_i )</th>
<th>( x_i )</th>
<th>( y_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 00000 0.70711</td>
<td>0.00000 0.01639</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 00000 0.41672</td>
<td>0.00000 0.04650</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 12486 1.82371</td>
<td>0.06640 0.08382</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 00000 0.23234</td>
<td>0.10770 0.11786</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 03136 0.83296</td>
<td>0.13535 0.15282</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 25274 1.47251</td>
<td>0.17059 0.18803</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 02648 3.08660</td>
<td>0.20643 0.22451</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 00000 0.12399</td>
<td>0.24960 0.26668</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 26442 0.49434</td>
<td>0.28000 0.29702</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 56775 0.79566</td>
<td>0.31778 0.33584</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 95864 1.11102</td>
<td>0.35625 0.37567</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.34443 1.57764</td>
<td>0.39545 0.41342</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.78764 2.17743</td>
<td>0.43540 0.45516</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 59722 3.01701</td>
<td>0.47611 0.49664</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 72421 4.42142</td>
<td>0.51762 0.53858</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 00000 0.06404</td>
<td>0.03596 0.05132</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 13226 0.20635</td>
<td>0.04033 0.05296</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 27326 0.34603</td>
<td>0.06472 0.08954</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 52426 0.50251</td>
<td>0.09521 0.11758</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 56967 0.67144</td>
<td>0.09939 0.10614</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 74322 0.85500</td>
<td>0.17953 0.20532</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 95548 1.05596</td>
<td>0.22337 0.25892</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 14497 1.28798</td>
<td>0.29282 0.30772</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 40199 1.52601</td>
<td>0.35322 0.38587</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 64448 1.80645</td>
<td>0.39846 0.41099</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 94991 2.13088</td>
<td>0.43778 0.45458</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 32214 2.51340</td>
<td>0.49208 0.51157</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 74494 2.90548</td>
<td>0.53780 0.55303</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 24056 3.50851</td>
<td>0.59034 0.60660</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 60074 4.42058</td>
<td>0.60434 0.62164</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.29449 5.83800</td>
<td>0.61961 0.63502</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 00000 0.03250</td>
<td>1.38661 1.41821</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 94165 0.09897</td>
<td>1.45077 1.48334</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 13445 0.16919</td>
<td>1.51434 1.55204</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 20507 0.24094</td>
<td>1.58524 1.61995</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 27806 0.31519</td>
<td>1.65582 1.69170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 35366 0.39217</td>
<td>1.72864 1.76524</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 43206 0.47199</td>
<td>1.80447 1.84294</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 53146 0.55490</td>
<td>1.88290 1.92384</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 59916 0.64135</td>
<td>1.96235 2.00587</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 66437 0.72139</td>
<td>2.04530 2.09227</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 77841 0.82542</td>
<td>2.12697 2.18325</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 87442 0.93380</td>
<td>2.20938 2.27661</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0 97541 1.05701</td>
<td>2.32658 2.37485</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 08123 1.19550</td>
<td>2.41766 2.48780</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.19267 1.29844</td>
<td>2.53334 2.59859</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.31024 1.37969</td>
<td>2.64376 2.70094</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.43478 1.49887</td>
<td>2.76183 2.81812</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.56709 1.62530</td>
<td>2.88952 2.95002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.70820 1.76110</td>
<td>3.03182 3.08467</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.85356 1.92766</td>
<td>3.10730 3.23230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 02216 2.10647</td>
<td>3.11055 3.23086</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 19851 2.29723</td>
<td>3.24741 3.35794</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 35990 2.49146</td>
<td>3.44975 3.57414</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 40256 2.71346</td>
<td>3.64270 3.74274</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 53777 2.94108</td>
<td>4.02588 4.14649</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 01055 3.24303</td>
<td>4.29113 4.41326</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.40514 3.56725</td>
<td>4.55384 4.69449</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.75076 3.95015</td>
<td>4.83464 5.01881</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 18397 4.41780</td>
<td>5.21034 5.40187</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 71828 5.01874</td>
<td>5.63584 5.84973</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 43991 5.86104</td>
<td>6.17040 6.47107</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 57245 7.28344</td>
<td>6.92550 7.31412</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b) Input Decision Thresholds and Output Levels of a Laplacian Quantizer

\[ x_i = \text{Input Threshold} \]
\[ y_i = \text{Output Level} \]
The Correlation Properties of the $v_i$ Sequence

The equations of representation of the first-order stationary sequence $\{x_i\}$ with zero mean and autocorrelation given by

$$E [x_i x_j] = \rho |i-j|$$  \hfill (1)

and given by

$$x_i = \alpha (x_{i+1} + x_{i-1}) + v_i \quad i = 1, \ldots, N$$  \hfill (2)

$$x_0 = \rho x_1 + v_0 ,$$  \hfill (3)

$$x_{N+1} = \rho x_N + v_{N+1}$$  \hfill (4)

where $\alpha = \rho/(1 + \rho^2)$. Multiplying (2) by $x_{i+k}$, taking expectation and using (1) we get

$$\rho |k| = \alpha (\rho |k-1| + \rho |k+1|) + E [v_i x_{i+k}]$$  \hfill (5)

which for $|k| > 1$ and $\alpha = \rho/(1+\rho^2)$, gives

$$E [v_i x_{i+k}] = 0 , \quad k \neq 0 , \quad 1 \leq i \leq N$$  \hfill (6)

Similarly for $k = 0$,

$$E [x_i v_i] = (1 - \rho^2)/(1 + \rho^2) \tilde{b}_2^2 ; \quad 1 \leq i \leq N .$$  \hfill (7)

Multiplication of (2) by $v_{i+k}$, taking expectations and use of (6) and (7) yields
\[ E[v_i v_{i+k}] = \beta^2 \left( \delta_{k,0} - \alpha \delta_{k,-1} - \alpha \delta_{k,1} \right) \quad 1 \leq i, i+k \leq N, \quad (8) \]

where \( \delta_{ij} \) is the Kronecker delta function. Similar procedure when applied to (3) gives.

\[ E[v_0 x_k] = (1 - \rho^2) \delta_{k,0}; \quad (9) \]

\[ E[v_0 v_k] = (1 - \rho^2) \delta_{k,0} - \alpha(1 - \rho^2) \delta_{k,1}; \quad (10) \]

\[ E[v_{n+1} x_k] = (1 - \rho^2) \delta_{k,N+1}; \quad (11) \]

\[ E[v_{N+1} v_k] = (1 - \rho^2) \delta_{k,N+1} - \alpha(1 - \rho^2) \delta_{k,N} \quad (12) \]
Main Processing and Storage Requirements of the Various Frequency-Domain Waveform Coders

1. Seven-Band Subband Coder (7/SBC/AQF)

1.1 Main Processing Requirements

(a) Analysis filter bank:

<table>
<thead>
<tr>
<th>QMF stage</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of additions (A1)</td>
<td>NT/2</td>
<td>NT/2</td>
<td>(\frac{7}{8}) NT/2</td>
<td>(\frac{23}{8}) NT/2/N samples</td>
</tr>
<tr>
<td>No. of multiplication (M1)</td>
<td>NT/2</td>
<td>NT/2</td>
<td>(\frac{7}{8}) NT/2</td>
<td>(\frac{23}{8}) NT/2/N samples</td>
</tr>
</tbody>
</table>

\(T = \) the length of the FIR filters used in the QMF filter bank

For \(N = 256\) and \(T = 32\)

No. of additions \(A_1 = 11776\)
No. of multiplications \(M_1 = 11776\)

(b) Quantizer Step-Size Estimation (every 16 msec):

No. of additions \(A_2 = 7 \times 31 = 217\)
No. of multiplications \(M_2 = 7 \times 32 = 224\)

(c) Normalization:

No. of division \(M_3 = 7 \times 32 = 224\)

. . . Total no. of additions/subtractions per sample

\[= \frac{(A_1 + A_2)}{256} = 47\]

Total no. of multiplications/divisions per sample

\[= \frac{(M_1 + M_2 + M_3)}{256} = 48\]
1.2 Memory size required for storing the intermediate variables:
(a) Filter bank: \( T + 2T + \frac{3}{2}T = \frac{13}{2} T = 208 \) words (for \( T = 32 \))
(b) Quantization: \( 7 \times 32 = 224 \) words
The size of RAM required = \( 208 + 224 = 430 \) words

1.3 If the second stage of the QMF filter-bank employs FIR filters
of length 24 and the third stage of length 20, the total no. of
additions/subtractions = 38/sample, the total no. of multiplications/
divisions = 39/sample, the size of RAM required = 370 words.

2. Fourteen-Band Subband Coder (14/SBC/AQF)

2.1 Main Processing Requirements
(a) Analysis filter bank:

<table>
<thead>
<tr>
<th>QMF stage</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of additions (A1)</td>
<td>NT/2</td>
<td>NT/2</td>
<td>( \frac{7}{8} (NT/2) )</td>
<td>( \frac{7}{8} (NT/2) )</td>
<td>( \frac{15}{4} (NT/2) )</td>
</tr>
<tr>
<td>No. of multiplications (M1)</td>
<td>NT/2</td>
<td>NT/2</td>
<td>( \frac{7}{8} (NT/2) )</td>
<td>( \frac{7}{8} (NT/2) )</td>
<td>( \frac{15}{4} (NT/2) )</td>
</tr>
</tbody>
</table>

\( T = \) the length of the FIR filters used in the QMF filter bank

For \( N = 256 \) and \( T = 32 \)

No. of additions \( A1 = 15360 \)
No. of multiplications \( M1 = 15360 \)

(b) Quantizer step-size Estimation (every 16 msec):

| No. of additions | \( A2 = 14 \times 15 = 210 \) |
| No. of multiplications | \( M2 = 14 \times 16 = 224 \) |
(c) Normalization:

No. of divisions $M_3 = 14 \times 16 = 224$

Total no. of additions/subtractions per sample

$= (A_1 + A_2)/256 = 61$

Total no. of multiplications/divisions per sample

$= (M_1 + M_2 + M_3)/256 = 62$

2.2 Memory size required for storing the intermediate variables:

(a) Filter bank: $T + 2T + \frac{7}{2}T + 7T = 432$ words ($T = 32$)

(b) Quantization: $16 \times 14 = 224$ words

The size of RAM required = 650 words

2.3 If the length of the FIR filters used in the 2nd, 3rd and 4th-stage of the QMF filter-banks are 24, 20, 16 respectively, the total no. of additions/subtractions = 44/sample, the total no. of multiplications/divisions = 45/sample, the size of RAM required = 500 words.

3. Seven-Band Complex Subband Coder (7/CSBC/AQF)

3.1 Main Processing Requirements

(a) Analysis filter bank:

<table>
<thead>
<tr>
<th></th>
<th>QMF 1st</th>
<th>2nd</th>
<th>3rd</th>
<th>CQMF 4th</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of additions (A1)</td>
<td>$\frac{NT}{2}$</td>
<td>$\frac{NT}{2}$</td>
<td>$\frac{7(NT)}{8}$</td>
<td>$\frac{14(NT)}{32}$</td>
<td>$\frac{15}{8}$ $NT$</td>
</tr>
<tr>
<td>No. of multiplications (M1)</td>
<td>$\frac{NT}{2}$</td>
<td>$\frac{NT}{2}$</td>
<td>$\frac{7(NT)}{8}$</td>
<td>$\frac{14(NT)}{32}$</td>
<td>$\frac{15}{8}$ $NT$</td>
</tr>
</tbody>
</table>

$T$ = the length of the FIR filters used in the filter-bank

For $N = 256$ and $T = 32$
No. of additions \( A_1 \) = 15360

No. of multiplications \( M_1 \) = 15360

(b) Rectangular to polar conversion:

No. of additions \( A_2 \) = 7 \times 16 = 112

No. of multiplications/divisions \( A_2 \) = 2 \times 7 \times 16 = 224

(c) Step-size estimation:

No. of additions \( A_3 \) = 7 \times 15 = 105

(d) Normalization:

No. of divisions \( M_3 \) = 7 \times 16 = 112

(e) Bit allocation using the VBA algorithm:

No. of subtractions \( A_4 \) = 7 \times 7 = 49

No. of multiplications \( M_4 \) = 7 \times 6 = 49

\[ . . . \text{Total no. of additions/subtractions per sample} \]
\[ = (A_1 + A_2 + A_3 + A_4)/256 = 62 \]

\[ \text{Total no. of multiplications/divisions per sample} \]
\[ = (M_1 + M_2 + M_3 + M_4)/256 = 63 \]

3.2 Memory size required for storing the intermediate variables:

(a) Filter bank: \( T + 2T + \frac{7T}{2} + 7T = 432 \) words (\( T = 32 \))

(b) Quantization: \( 16 \times 14 = 224 \) words

\[ . . . \text{The size of RAM required} = 650 \text{ words} \]
4. Large blocksize (128) ATC/ABA Coder

4.1 Main Processing Requirements

(a) Transformation using the fast algorithm proposed by Chen

\[
\text{No. of } +/-(A1) = \frac{3N}{2} (\log_2 N - 1) + 2 = 1154
\]

\[
\text{No. of } \times/\div (M1) = N \log_2 N - \frac{3N}{2} + 4 = 708, \text{ for } N = 128.
\]

(b) Basis spectrum estimation:

\[
\text{No. of additions (A2)} = 7 \times 16 = 112
\]

\[
\text{No. of multiplications (M2)} = 128
\]

(c) Linear interpolation of the 16 primary coefficients:

\[
\begin{align*}
\text{No. of subtractions} & = 7 \\
\text{No. of divisions (M3)} & = 7 \\
\text{No. of additions} & = 7.15 = 105 \\
\end{align*}
\]

\[\therefore \text{No. of } +/(A3) = (7 + 105)/2 = 56 \text{ for 1 frame}\]

(d) Bit allocation:

From the equation

\[
R_i = \frac{1}{N} + \frac{1}{2} \log_2 \hat{g}_i^2 - \frac{1}{2N} \sum_{j=1}^{N} \log_2 \hat{g}_j^2
\]

\[\text{2 additions/sample} \quad \text{127 additions}\]

\[
\text{No. of } +/-(A4) = 2 \times 128 + 127 = 383/2 \text{ frames) = 192/frame}
\]
(e) Normalization:

No. of multiplication/divisions (M4) = 256

Total no. of additions/subtractions per sample

\[ (A_1 + A_2 + A_3 + A_4) \times 128 = 12 \]

Total no. of multiplications/divisions per sample

\[ (M_1 + M_2 + M_3 + M_4) \times 128 \approx 9 \]

4.2 Memory size required for storing the intermediate variables:

Buffers for:

(a) 2 frames of input data: 256 words
(b) 1 frame of normalized data: 128 words
(c) 2 frames of transform coefficients: 256 words
(d) 1 frame of estimated basis spectrum: 128 words
(e) 1 frame of normalized transform coefficients: 128 words

The size of RAM required = 900 words

5. Small blocksize(9) PST/ABA Coder

5.1 Main Processing Requirements

(a) Transformation using fast algorithm:

No. of +/- (A1) = 24 x 30 = 720 \{ For 24 vectors
No. of x/\div (M1) = 24 x 22 = 528 \} in 216 samples

(b) Linear interpolation using the 2 end-elements:

No. of +/- (A2) = 24x8 = 192 \{ For 216
No. of x/\div (M2) = 24 \} samples

(c) Residue calculation:

No. of +/- (A3) = 24x8 = 192
(d) Normalization:
No. of +/- (A4) = 215
No. of x/* (M3) = 216 + 216 = 432
(e) Basis spectrum estimation, bit allocation and quantization:
No. of +/- (A5) = 87
No. of +/- (M4) = 216
.'. Total no. of additions/subtractions per sample
= (A1 + A2 + A3 + A4 + A5)/216 = 7
Total no. of multiplications/divisions per sample
= (M1 + M2 + M3 + M4)/216 = 6

5.2 Memory size required for storing the intermediate variables:
Buffer for
(a) 216 input samples
(b) 1 block of 8 transform coefficient vectors: 64
(c) estimated basis spectrum : 8
.': The size of RAM required = 300
REFERENCES


118. NEC Electronics (Europe), NEC µPD 7720, User's Manual.


121. CCITT, Study Group XVIII, Swiss PTT, "Simulations wideband speech and data transmission on 64 kbit/s", Contribution No. 204, Sept. 1983.


149. CCITT, Study Group XVIII, Italian PTT, "Coding of audio signals with a band larger than 4 kHz", Delayed Document EK, 1982.


