Enhancement of band-limited speech signals

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

Additional Information:


Metadata Record: [https://dspace.lboro.ac.uk/2134/20947](https://dspace.lboro.ac.uk/2134/20947)

Publisher: © P.J. Patrick

Rights: This work is made available according to the conditions of the Creative Commons Attribution-NonCommercial-NoDerivatives 2.5 Generic (CC BY-NC-ND 2.5) licence. Full details of this licence are available at: [http://creativecommons.org/licenses/by-nc-nd/2.5/](http://creativecommons.org/licenses/by-nc-nd/2.5/)

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.

![Creative Commons Licence](https://creativecommons.org/licenses/by-nc-nd/2.5/)

**Attribution-NonCommercial-NoDerivs 2.5**

You are free:
- to copy, distribute, display, and perform the work

Under the following conditions:

**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)](http://creativecommons.org/licenses/by-nc-nd/2.5/)

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
ENHANCEMENT OF BANDLIMITED
SPEECH SIGNALS

BY

P.J. PATRICK

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

Supervisor: Dr. C.S. Xydeas.
Loughborough University of Technology
Department of Electronic and Electrical Engineering

© by P.J. Patrick, 1983.
ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Dr C S Xydeas whom I am privileged to have had as my research supervisor. His enthusiasm in the field of digital signal processing has been a strong source of encouragement to me. His technical guidance and valuable criticisms have greatly influenced the course of this research. I am also thankful to my former supervisor Professor Raymond Steele (who is now at Southampton University) as he enabled me to commence my research into speech signal processing. I am extremely grateful to him for his continued technical support and assistance in producing some of the publications during his period of work at Bell Telephone Laboratories, USA. My thanks as also extended to my interim supervisor Mr R J Clarke of the Department of Electronic and Electrical Engineering at Loughborough University of Technology.

I also wish to thank the Science and Engineering Research Council for providing the financial support under the CASE award scheme with British Telecom Research Laboratories (BTRL) at Martlesham Heath, Ipswich, Suffolk. I wish to thank all those people who organised the CASE award scheme for me, they are: J W R Griffiths, Professor of Electronics in the Department of Electronic and Electrical Engineering at Loughborough University of Technology; Mr C J Hughes, Deputy Director at BTRL, Professor Raymond Steele; Mr Ian Groves from the Human Factors Department at BTRL; Mr C South and Dr O Guard also from BTRL. I am grateful to the people at BTRL for enabling me to carry out some of the hardware implementation associated with my research work.
I would also like to thank Professor I R Smith, Head of Electronic and Electrical Engineering Department at Loughborough University of Technology for providing all of the research facilities. I am thankful to the staff and student members of this Department whose assistance was valuable to my work. I am especially grateful to Dr A P Clark, Mr A F Erwood and Dr C C Evei for their helpful comments and suggestions and to Mr S N Koh, Mr W K Cham and Mr V Tse for their collaborative work with my research. I am also extremely grateful to Dr J N Holmes of JSRU Cheltenham, Glos. for all of his technical comments and suggestions.

Thanks are also due to the Computer Centre Staff at Loughborough University of Technology for running my lengthy batch-jobs on the ICL 1900 computer and to the extensive help from Mr G P Gerrard and the Program Advisory Service.

My gratitude is also extended to Mrs Janet Redman of the Department of Mechanical Engineering at Loughborough University of Technology for drawing some of the diagrams included in this thesis and to Pauline and Liz for their forbearance in typing this thesis on the word processor. I would also like to thank Mrs B Wright for typing the unwieldy equations, symbols and figure captions.

Finally I express my deepest gratitude to my parents for their moral support whose confidence in me has made my study possible.

To them I dedicate this thesis.
Ph.D. Synopsis

**Enhancement of Band Limited Speech Signals**

Using speech transmission via a telephone channel of bandwidth 0.3 to 3.4 kHz, it is desired to gain a wider subjective bandwidth at the receiver output than that afforded by the channel itself. The channel is considered here to be a realisable perfect band pass filter, i.e. free of noise or dispersion.

Two distinct approaches are adopted to meet this objective:-

Firstly, a 'receiver-only' processing type of system where the bulk of the spectral enhancement procedure occurs at the receiver. This would be particularly well suited to radio and television 'phone-in' programmes. The speech signal originates as a 0.3 to 3.4 kHz band limited version where it is left to the receiver to regenerate spectral components outside this range.

A novel method is developed to create this situation by feeding the 0.3 to 3.4 kHz signal into a non-linear transfer function in order to regenerate components outside this band range. This regenerated signal is added to the existing band limited speech to provide a subjectively wider band signal.
The second approach is a 'transmitter-receiver' bandwidth compression system in which the speech originates as a 0.3 to 7.6 kHz signal. The signal is spectrally compressed at the transmitter to yield a band limited version of 0.3 to 3.4 kHz to be conveyed via the 'telephone-channel'. The receiver applies spectral expansion in order to recover as accurately as possible the original bandwidth speech signal. This latter approach is a more useful adjunct to audio teleconferencing systems where wide band, good quality speech is available at the input to the system.

Three methods developed are encompassed by this second case. The first two methods are based upon spectral compression in the discrete frequency domain. Both techniques attempt to preserve those frequency coefficients which hold the greatest significance as far as the spectrum is concerned. The first method is termed 'frequency-mapping', the second is employed as a 'random-phase substitution' scheme.

Lastly in this case, a time domain process called 'voiced/unvoiced band switching system' is developed and utilised by filtering the wide band signal into two 3 kHz-bands and transmitting the band with greatest energy.

The frequency mapping and the voiced/unvoiced band switching techniques play an important role in assisting digital speech transmission systems whereby the speech signal is pre-processed prior to efficient encoding by APCM, ADPCM, ADM and CVSD. As a result a subjectively improved decoded signal for a given bit rate is obtained.
Also included in this project are notions which are briefly examined
to discern voiced speech from unvoiced signals as this facility is
required by all of the spectral compression algorithms presented in
this thesis.

The experimentation is performed by computer simulation using
informal subjective listening criteria to evaluate the results.
These indicate a definite improvement over normal telephonic
bandwidth speech for all the systems examined.
Glossary of Principal Symbols and Nomenclature

ADM - Adaptive Delta Modulation
ARMAP - Adaptive Frequency Mapping
agc - Automatic Gain Control
AM - Amplitude Modulation
\( a_n \) - Filter tap coefficients (feed-forward stage)
APC - Adaptive Predictive Coding
APCM - Adaptive Pulse Code Modulation
ATC - Adaptive Transform Coding
B - Number of bits per quantized sample
ber - bit error rate
bin - Two adjacent frequency samples in the DFT field, separated by \( \Delta f = F_s/N \), are \( 1 \) bin apart
\( b_n \) - Filter tap coefficients (feed-back stage)
BPF - Bandpass filter
BW - Maximum bandwidth of the input signal \( x(t) \)
\( C_i \) - Signal autocorrelation
CFDM - Constant Factor Delta Modulation
Ch - Channel (analogue or digital)
Codec - Coder/decoder configuration
CVSD - Continuous Variable Slope Delta modulation
dc - direct current, equivalent to long term average level
DCT - Discrete Cosine Transform
DEMUX - Demultiplexing
DFT - Discrete Fourier Transform
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>Delta Modulation</td>
</tr>
<tr>
<td>DPCM</td>
<td>Differential Pulse Code Modulation</td>
</tr>
<tr>
<td>DSTFT</td>
<td>Discrete Short Time Fourier Transform</td>
</tr>
<tr>
<td>E(.)</td>
<td>Ensemble average of '.', equal to the expected value '.'</td>
</tr>
<tr>
<td>{e_i}</td>
<td>error sequence</td>
</tr>
<tr>
<td>f</td>
<td>frequency (Hertz)</td>
</tr>
<tr>
<td>FDM</td>
<td>Frequency Division Multiplexing</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite impulse response</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>FM</td>
<td>Frequency Modulation</td>
</tr>
<tr>
<td>FMAP</td>
<td>Frequency Mapping</td>
</tr>
<tr>
<td>f_{max}</td>
<td>One half the Nyquist frequency</td>
</tr>
<tr>
<td>F_s</td>
<td>Sampling frequency</td>
</tr>
<tr>
<td>FT</td>
<td>Fourier Transform</td>
</tr>
<tr>
<td>F(z)</td>
<td>Filter transfer function</td>
</tr>
<tr>
<td>Fo</td>
<td>Pitch frequency of voiced speech</td>
</tr>
<tr>
<td>G</td>
<td>Fixed amplifier gain value</td>
</tr>
<tr>
<td>h_f</td>
<td>high frequency</td>
</tr>
<tr>
<td>HFR</td>
<td>High frequency regeneration</td>
</tr>
<tr>
<td>h(n)</td>
<td>Filter impulse response (also used as an analysis window for the short time Fourier Transform)</td>
</tr>
<tr>
<td>HPF</td>
<td>highpass filter</td>
</tr>
<tr>
<td>Hz</td>
<td>Hertz = 1 cycle per second</td>
</tr>
<tr>
<td>H(z)</td>
<td>Filter transfer function</td>
</tr>
<tr>
<td>i</td>
<td>i^{th} sampling instant (discrete time index)</td>
</tr>
<tr>
<td>IIR</td>
<td>Infinite impulse response</td>
</tr>
<tr>
<td>Im(.)</td>
<td>Imaginary part of '.'</td>
</tr>
<tr>
<td>INT(.)</td>
<td>Rounded down integer value of '.'</td>
</tr>
</tbody>
</table>
\( j \) - \( \sqrt{-1} \), imaginary unit vector
\( k \) - \( k^{th} \) frequency ordinate
kernel - \( \sin (x)/x \) function
KLT - Karhunen Loeve Transform
kHz - Kilohertz = 1000 Hertz
LDM - Linear Delta Modulation
lf - low frequency
LDM - Linear Delta Modulation
LPC - Linear Prediction Coding
LPF - Lowpass filter
L(t) - binary waveform
MUX - Multiplexing
\( N \) - Number of samples in the DFT block (or any other orthogonal transform block)
n.i. - near-instantaneous companding
NINT(.) - Nearest integer value of '.
\( N^2 \) - mean square quantization noise
\( p(.) \) - probability distribution of '.
PCM - Pulse Code Modulation
pdf - probability distribution function
PICOR - Pilot Controlled Overtone Reproduction
PSTN - Public Switched Telephone Network
\( P(z) \) - Predictor function
\( P_0 \) - Pitch period
\( Q \) - Quantizer
QMF - Quadratic Mirror Filtering
\( q_i \) - Quantizer decision levels
\( \text{Re}(.) \) - Real part of '.
Rx - Receiver terminal
(vii)

\( s \) - Laplacian variable = \( \sigma + j \omega \)

SBC - Sub-band coding

SDM - Space Division Multiplexing

SNR - Signal to Noise Ratio

SP-SNR - Spectral SNR

SP-SNRSEG - Segmented Spectral SNR

SSBSC - Single Sideband Suppressed Carrier Modulation

STFT - Short Time Fourier Transform

T - Sampling interval

TBR - Transmission bit rate

TDM - Time Division Multiplexing

\( t_L \) - Record length of data used for the DFT

Tx - Transmitter terminal

T(z) - Filter transfer function

u(t) - Unvoiced speech time function

UV - Unvoiced speech

V - Voiced speech

v(t) - Voiced speech time function

VUBS - Voiced/unvoiced band switching system

V/UV - Voiced/unvoiced speech detector

V/UV/S - Voiced/unvoiced/silence discriminator

V(z) - Vocal tract response

w(n) - Window function applied prior to DFT analysis

\( W_N \) - \( \exp(-j2\pi/n) \) DFT phasor

\( \{x_i\} \) - samples of \( x(t) \)

\( \hat{x}_i \) - estimate of \( x_i \)

\( X(j\omega) \) - Fourier Transform of \( x(t) \)

\( \{X_k\} \) - DFT sequence of \( x_i \)
|\{X_c(k)\}| - DCT sequence of x_i

x(t) - input analogue signal, sometimes abbreviated as 'x'

\hat{x}(t) - reconstructed analogue signal

\tilde{x}(t) - Hilbert Transform of x(t)

X(\omega_n,t) - Short-time Fourier Transform of x(t) evaluated at frequency \omega_n

y(t) - time waveform of a spectrally compressed signal

z - exp (sT)

z^{-1} - delay of 1 sampling period

ZC - zero crossing (or waveform axis crossing)

\alpha_m - Position of zeros of an FIR or IIR filter

\beta_m - Position of poles of an IIR filter

\Delta - quantizer step size

\Delta f - frequency resolution of the DFT sequence = N/F_s

\rho_m - Normalised autocorrelation coefficient function corresponding to m sample shifts

\sigma_x - rms value of input signal x(t)

\sigma_x^2 - variance of input signal x(t)

\theta_k - phase of X_k

\omega - angular frequency = 2\pi f

\langle . \rangle - time average of '.'

\langle . \rangle^* - complex conjugate of '.'

x(t)→X(jw) - x(t) and X(jw) are Fourier transform pairs

a(t)*b(t) - convolution of a(t) and b(t)
## Volume 1

### Chapter 1

**Enhancement of Bandlimited Speech Signals**

#### ORGANISATION OF THESIS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Structure of the Thesis</td>
<td>6</td>
</tr>
<tr>
<td>1.3 Summary of Main Results</td>
<td>11</td>
</tr>
</tbody>
</table>

### Chapter 2

**Speech Communications**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Speech Production and Perception</td>
<td>15</td>
</tr>
<tr>
<td>2.1.1 Speech Production</td>
<td>15</td>
</tr>
<tr>
<td>2.1.2 The Ear and Hearing</td>
<td>21</td>
</tr>
<tr>
<td>A The Outer Ear</td>
<td>21</td>
</tr>
<tr>
<td>B The Middle Ear</td>
<td>22</td>
</tr>
<tr>
<td>C The Inner Ear</td>
<td>22</td>
</tr>
<tr>
<td>2.1.3 Loudness</td>
<td>25</td>
</tr>
<tr>
<td>2.2 Characterisation of Speech Waveforms</td>
<td>26</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>2.3</td>
<td>Voice Communications</td>
</tr>
<tr>
<td>2.3.1</td>
<td>The Telephone System</td>
</tr>
<tr>
<td>2.3.1.1</td>
<td>The Telephone Set</td>
</tr>
<tr>
<td>2.3.1.2</td>
<td>The Subscriber's Line and Local Exchange</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Causes of Degradation of the Speech Signal by the Telephone System</td>
</tr>
<tr>
<td></td>
<td>A. The Talker</td>
</tr>
<tr>
<td></td>
<td>B. The Talker's Environment</td>
</tr>
<tr>
<td></td>
<td>C. The Transducer</td>
</tr>
<tr>
<td>2.3.2.1</td>
<td>Noise Introduced by the Telephone Network</td>
</tr>
<tr>
<td></td>
<td>A. Crosstalk</td>
</tr>
<tr>
<td>2.3.2.2</td>
<td>Amplitude/Frequency and Group Delay Characteristics</td>
</tr>
<tr>
<td></td>
<td>A. The Microphone</td>
</tr>
<tr>
<td></td>
<td>B. Transmission</td>
</tr>
<tr>
<td></td>
<td>C. Mis-match</td>
</tr>
<tr>
<td></td>
<td>D. Delay</td>
</tr>
<tr>
<td></td>
<td>E. Non-linearity</td>
</tr>
</tbody>
</table>
2.3.2.3 Time Varying Effects

A. Speech Operated Devices
B. Echo Suppression
C. Time Assignment Speech Interpolation (TASI)
D. Loud Speaking Telephones

2.3.2.4 Random Effects

A. Fading Radio Channels
B. Frequency Offset
C. Phase Jitter and Jumps

2.3.3 Audio Teleconferencing

2.3.3.1 General
2.3.3.2 Types of Teleconferencing Available Hitherto
2.3.3.3 Fundamental Requirements for Audio Teleconferencing
A. Loudness
B. Bandwidth and Speech quality
C. Howl Prevention

2.3.3.4 'Orator' - The British Telecom Audio Teleconferencing System
2.3.3.5 The Plessey Remote Conference System

2.3.4 'Phone-in' Contributions to Broadcast Radio and TV Programmes.
### Chapter 3: A REVIEW OF BANDWIDTH REDUCTION TECHNIQUES APPLIED TO SPEECH SIGNALS

#### 3.1 Introduction

Page 60

#### 3.2 Waveform Coders

Page 62

##### 3.2.1 Pulse Code Modulation

Page 63

##### 3.2.1.1 Time Invariant Quantizers

Page 64

##### 3.2.1.2 Adaptive Quantization

Page 68

##### 3.2.1.3 Dithered Quantization

Page 69

##### 3.2.2 Differentially Coding Systems

Page 70

##### 3.2.2.1 Differential Pulse Code Modulation

Page 72

##### 3.2.2.2 Delta Modulation

Page 74

- A. Adaptive Delta Modulation
- B. Constant Factor Delta Modulation
- C. Continuously Variable Slope Delta Modulation

Page 76

#### 3.3 Analysis - Synthesis Systems

Page 82

##### 3.3.1 Channel Vocoder

Page 83
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.3.1.1</td>
<td>Redundancy Reduction with Channel Vcoders</td>
<td>87</td>
</tr>
<tr>
<td>A.</td>
<td>Peak-picking</td>
<td>87</td>
</tr>
<tr>
<td>B.</td>
<td>Linear Transformation of the Channel Signal</td>
<td>87</td>
</tr>
<tr>
<td>C.</td>
<td>Pattern Matching Vcoders</td>
<td>88</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Correlation Vcoders</td>
<td>89</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Formant Vocoder</td>
<td>91</td>
</tr>
<tr>
<td>3.3.4</td>
<td>Orthogonal Vocoder</td>
<td>94</td>
</tr>
<tr>
<td>3.3.5</td>
<td>Homomorphic Vocoder</td>
<td>96</td>
</tr>
<tr>
<td>3.3.6</td>
<td>Maximum Likelihood Vocoder</td>
<td>101</td>
</tr>
<tr>
<td>3.3.7</td>
<td>LPC Vocoder</td>
<td>104</td>
</tr>
<tr>
<td>3.3.8</td>
<td>Articulatory Vocoder</td>
<td>106</td>
</tr>
<tr>
<td>3.3.9</td>
<td>Frequency Dividing Vocoder</td>
<td>107</td>
</tr>
<tr>
<td>3.3.9.1</td>
<td>Vobanc</td>
<td>109</td>
</tr>
<tr>
<td>3.3.10</td>
<td>Analytic Rooter</td>
<td>109</td>
</tr>
<tr>
<td>3.3.11</td>
<td>Phase Vocoder</td>
<td>115</td>
</tr>
</tbody>
</table>
3.4 Intermediate Systems

3.4.1 Frequency Domain Coding

3.4.1.1 Voice Excited Vocoder

3.4.1.2 Sub-band Coding

3.4.1.3 Adaptive Transform Coding

A. Block Transform
B. Quantization Strategy
C. Adaptation Strategy
D. Noise Shaping

3.4.1.4 Discrete Fourier Transform Methods

A. Discrete Short-time Fourier Transform

3.4.1.5 Frequency Coding based on the Phase Spectrum

A. Co-phase Processing
B. Phase-quantizing
C. Pitch Synchronous Processing
D. Continuous phase Processing

3.4.2 Time Domain Coding
3.4.2.1 Adaptive Predictive Coding

A. Prediction based on Spectral Fine Structure
B. Prediction based on Short Time Spectral Envelope
C. Combining the Two Types of Prediction
D. Perceptual Criteria for Optimizing Predictive Coding

3.4.2.2 Time Domain Harmonic Scaling

3.4.2.3 Time Encoded Speech

3.4.3 Comparison of Speech Coders

3.4.3.1 Transmission-Rate Profile
3.4.3.2 Characteristics of the Coders

3.5 Wide Band Speech Processing

3.5.1 Voice Excited Vocoder

3.5.1.1 The Spectral Flattener

3.5.2 Digital Compression of Wideband Speech Signals

3.5.2.1 Coding the Zero Crossings of the Upper Band
3.5.2.2 Sub-Nyquist Sampling
3.5.2.3 Pitch Halving
3.5.2.4 Variable Sampling Rate with Constant Bit rate
3.5.2.5 Sub band Coding of Wideband Speech
| 3.5.3 | Quality Improvement of Telephone Speech | 169 |
| 3.5.3.1 | Amplitude/Frequency Equalisation | 169 |
| 3.5.3.2 | Extension of Bandwidth | 170 |
| A. | Low Frequency Synthesis | 170 |
| B. | High Frequency Synthesis | 171 |
| 3.5.3.3 | Noise Reduction System | 175 |
| 3.6 | Voiced/Unvoiced Detection | 177 |
| 3.6.1 | Telephone Quality Speech | 177 |
| 3.6.1.1 | Decision Based on LDM and Zero Crossing Rates | 179 |
| 3.6.2 | Wideband Speech Analysis | 181 |
| 3.6.3 | Voiced/Unvoiced Decision Using Context | 183 |
4.5 Detection based upon Zero Crossing Rate of the Time Waveform

4.6 Detection based upon Power Level Measurement

4.7 Detection of Voiced Speech based upon the First Shift Autocorrelation Coefficient

4.8 Discussion

Volume 2
Chapter 5 SPEECH BANDWIDTH ENHANCEMENT

5.1 Introduction

5.2 High Frequency Regeneration (HFR) by Spectral Duplication

5.2.1 The Spectral Duplication Process

5.3 HFR by the Synthesis of an Out-of-Band formant for Voiced Speech

5.3.1 Experimental Procedure

5.3.2 HFR Excitation

5.3.3 The Resonant Network

5.3.4 Results
5.4 HFR for Unvoiced Speech

5.4.1 Development of the High Frequency Regenerator for the Unvoiced Speech Sounds

5.4.1.1 Bandpass Filtered Noise

5.4.1.2 Use of a 'W' Transfer Function for Unvoiced Speech Signals

5.4.1.3 Cubic Transfer Function

5.4.1.4 Hybrid Regenerator

5.4.1.5 'W' Transfer Function Applied only for /s/ Sounds

5.4.1.6 Results

5.4.2 Transmission of the Waveform Envelope for /s/ Sounds

5.4.3 Shaping the Spectrum of the Processed /s/ Signals

5.5 Digital Techniques in Conjunction with Bandwidth Enhancement Post-processing

5.5.1 Optimization of APCM

5.5.2 Optimization of DPCM

5.5.3 Adaptive Delta Modulation (ADM)

5.5.3.1 Constant Factor Delta Modulation (CFDM)

5.5.3.2 Continuously Variable Slope Delta Modulation (CVSD)
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5.4</td>
<td>Speech Enhancement System incorporating the Digital Codecs</td>
<td>253</td>
</tr>
<tr>
<td>5.5.4.1</td>
<td>Results of using the Digital Codecs alone</td>
<td>254</td>
</tr>
<tr>
<td>5.5.4.2</td>
<td>Addition of HFR</td>
<td>257</td>
</tr>
<tr>
<td>5.5.5</td>
<td>Future recommendations of HFR</td>
<td>258</td>
</tr>
<tr>
<td>5.6</td>
<td>Base-band Synthesis</td>
<td>259</td>
</tr>
<tr>
<td>5.6.1</td>
<td>Low-frequency Synthesis by a Local Oscillator driven at Pitch Frequency</td>
<td>260</td>
</tr>
<tr>
<td>5.6.1.1</td>
<td>Sliding Block Strategy</td>
<td>262</td>
</tr>
<tr>
<td>5.6.1.2</td>
<td>Pitch Detection by the Cepstrum Method</td>
<td>265</td>
</tr>
<tr>
<td>5.6.1.3</td>
<td>Pitch Detection by Data Reduction</td>
<td>268</td>
</tr>
<tr>
<td></td>
<td>A. Data Construction and Reduction</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.6.1.4 Results</td>
<td>270</td>
</tr>
<tr>
<td>5.6.2</td>
<td>Inclusion of the 2nd Harmonic plus Spectral Shaping</td>
<td>272</td>
</tr>
<tr>
<td>5.6.2.1</td>
<td>Vocal Tract Frequency Response</td>
<td>273</td>
</tr>
<tr>
<td>5.6.2.2</td>
<td>Results</td>
<td>274</td>
</tr>
<tr>
<td>5.7</td>
<td>Discussion, Conclusion and Recommendations</td>
<td>277</td>
</tr>
<tr>
<td>5.8</td>
<td>Note on Publication</td>
<td>283</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>6.1</td>
<td>Introduction</td>
<td>284</td>
</tr>
<tr>
<td>6.2</td>
<td>The General Compression System</td>
<td>286</td>
</tr>
<tr>
<td>6.3</td>
<td>The Voiced/Unvoiced Band Switching System</td>
<td>289</td>
</tr>
<tr>
<td>6.3.1</td>
<td>Results</td>
<td>291</td>
</tr>
<tr>
<td>6.4</td>
<td>Bandwidth Compression Techniques in the Frequency Domain</td>
<td>293</td>
</tr>
<tr>
<td>6.4.1</td>
<td>The use of Windows for Harmonic Analysis with the DFT</td>
<td>295</td>
</tr>
<tr>
<td>6.4.1.1</td>
<td>Time Sampling</td>
<td>296</td>
</tr>
<tr>
<td>6.4.1.2</td>
<td>Time Limitation</td>
<td>296</td>
</tr>
<tr>
<td>6.4.1.3</td>
<td>Time Convolution</td>
<td>297</td>
</tr>
<tr>
<td>6.4.1.4</td>
<td>Time Weighting Functions</td>
<td>298</td>
</tr>
<tr>
<td>6.4.1.5</td>
<td>Harmonic Analysis</td>
<td>303</td>
</tr>
<tr>
<td>6.4.1.6</td>
<td>Application of Windowing for Bandwidth Compression</td>
<td>307</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Bandwidth Compression using the Second Order Spectrum</td>
<td>309</td>
</tr>
</tbody>
</table>
6.4.3 Spectral Compression by Frequency Mapping

6.4.3.1 Frequency De-mapping
6.4.3.2 Frequency Mapping System
6.4.3.3 Compression Law
6.4.3.4 Results and Conclusions
6.4.3.5 Adaptive Frequency Mapping

A. The Mapping Laws
B. When to use Adaptive Frequency Mapping
C. Selecting the Mapping Law
D. Results and Discussion

6.4.4 Bandwidth Compression in the Frequency Domain using an Artificial Phase Function

6.4.4.1 Operation of the Phase Processor

A. The Transmitter
B. The Receiver
C. The Channel

6.4.4.2 Results

6.4.4.3 Application of the Random Phase Processor to /s/ Sounds
6.4.4.4 Results

A. Time and Spectral Plots
B. Informal Subjective Listening Tests
C. Segmented Spectral SNR

6.4.4.5 Discussion and Conclusion

6.5 Time Domain Harmonic Scaling of Unvoiced Speech Signals

6.5.1 Time Domain Harmonic Compression

6.5.2 Time Domain Harmonic Expansion

6.5.3 Results

6.5.3.1 Comparison by Informal Listening Tests

6.5.4 Discussion and Conclusion

6.6 Digital Coding of the 300 to 3400Hz Speech Channel

6.6.1 Sampling Frequency Conversion

6.6.2 Results and Discussion

6.6.2.1 Informal Subjective Testing

6.6.2.2 Waveforms
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall Discussion and Conclusions</td>
<td>391</td>
</tr>
<tr>
<td>Note on Publications</td>
<td>394</td>
</tr>
</tbody>
</table>

**Chapter 7  RECAPITULATION  396 - 422**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>396</td>
</tr>
<tr>
<td>The Spectral Extension Techniques</td>
<td>398</td>
</tr>
<tr>
<td>Spectral Compression Techniques</td>
<td>401</td>
</tr>
<tr>
<td>Methods of Appraisal</td>
<td>408</td>
</tr>
<tr>
<td>Spectral - SNR Measurements</td>
<td>408</td>
</tr>
<tr>
<td>Informal Subjective Listening Tests</td>
<td>411</td>
</tr>
<tr>
<td>Spectrographic Displays</td>
<td>414</td>
</tr>
<tr>
<td>Closing Remarks</td>
<td>422</td>
</tr>
</tbody>
</table>
Appendices

A.1 Digital Filtering

A.1.1 Recursive filter design

A.1.1.1 Frequency transformations
A.1.1.2 Gain normalisation

A.1.2 Non-recursive filter design

A.1.2.1 Design by the window method
A.1.2.2 Design by the sampling method

A. The minimisation algorithm
B. Bandpass filters
C. Programming considerations

A.1.2.3 Sectioned convolutions

A.2 High Frequency Regeneration

A.2.1 Logarithmic compression
A.2.2 The W-transfer function

A.3 Spectral Peak Synthesis
A.4 Fortran Program Listings

List 1 - Subroutine 'BUTTER' used to calculate the poles and zeros of a Butterworth response lowpass filter digital filter

List 2 - Subroutine 'BPASS' used to calculate the poles and zeros of a bandpass filter from the poles and zeros of a lowpass prototype

List 3 - Subroutine 'COEFF' used to compute the coefficients in a serial form realisation of a digital filter

List 4 - Subroutine 'FREQ' used to compute the gain and phase characteristics of a digital filter from the coefficients of its second order section realisation

List 5 - Subroutine 'SETUP' used to design a digital bandpass filter derived from a lowpass Butterworth response prototype using a combination of subroutines BUTTER, BPASS, COEFF and FREQ

List 6 - Function 'HIIR' used to implement the bandpass filter designed by a call to 'SETUP' subroutine
<table>
<thead>
<tr>
<th>List</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Subroutine &quot;FILTER&quot; used to compute the impulse response of an FIR filter</td>
<td>457</td>
</tr>
<tr>
<td>8</td>
<td>Subroutine 'TR' which contains the transition values for the lowpass filter set up by a call to 'FILTER'</td>
<td>459</td>
</tr>
<tr>
<td>9</td>
<td>Subroutine 'INTERP' used to calculate the interpolated values of the transition values not specified by subroutine 'TR'</td>
<td>462</td>
</tr>
<tr>
<td>10</td>
<td>Subroutine 'FCONV' used to perform segmented convolution</td>
<td>463</td>
</tr>
<tr>
<td>11</td>
<td>Subroutine 'LOGCOM' used to apply logarithmic compression</td>
<td>464</td>
</tr>
<tr>
<td>12</td>
<td>Subroutine 'W' used to apply a W-transfer function</td>
<td>465</td>
</tr>
<tr>
<td>13</td>
<td>Subroutine 'FORMANT' used to synthesise an additional formant to the speech signal</td>
<td>466</td>
</tr>
<tr>
<td>14</td>
<td>Subroutine 'RESPONSE' used to compute the gain and phase response of the resonant filter in 'FORMANT' subroutine.</td>
<td>468</td>
</tr>
</tbody>
</table>

References 469 - 484
To many people it may seem novel to consider speech as the topic of scientific or engineering study. One reason is that it is so much a part of everyday taken-for-granted behaviour that we seldom consider the need for its description or explanation. Speech is always available since it is produced by the human body without any tool; it can be varied from a soft whisper to a loud shout; it fills the entire space around the speaker, goes around obstacles, and thus does not require a direct line of connection with the hearer. It does not depend on light as do optical signals and so can be used day or night; it leaves the body almost entirely free for other activities and requires but little energy. Speech is not merely tongue-wagging, larynx buzzing and listening. It is much more the result of the brain doing its job as a manager of muscles to keep you going in your situation. Similarly, it would be misleading to use the word 'listen' in describing the functions of the ears in everyday speech. We do not 'switch-on' our ears just to catch a few seconds of speech. Our ears are actively interested in what is going on.

The knowledge which enables man to associate sounds with meanings is called language – this knowledge is the art of identifying sounds and the rules which associate sounds and patterns of sounds with meanings. Our performance involving this knowledge is speech. (ref 1)
When we describe speech we go beyond the description of language in the attempt to account for all the additional factors incorporated in language performance e.g. the physical nature of sounds and the human capability for articulating and receiving different speech sounds. Speech, then, is the complex combination of many factors, and these are:

(i) the acoustics of speech;
(ii) the sounds of language (phonology);
(iii) descriptions of language;
(iv) the psychological view of speech, and
(v) the sociological view of speech.

The Acoustics of Speech. As speech exists outside of the human organism, it is a form of energy or sound waves. It is not difficult to study this aspect of speech so long as we have the capabilities for understanding acoustics i.e. the science of sounds. Acoustics provides us with the perspective on the physical nature of speech.

Phonology. Only some portions of the acoustic patterns produced by the vocal mechanism are relevant to the basic sound patterns of a language. We need to understand the distinctive features of sounds as a strategy for classifying the sounds and to differentiate between them.

Language. As we have already said, speech is the performance of language. A language relates sound and meaning by specified rules.
The Psychological View. This considers the question of why and how should the speaker-listener use language. Vastly differing theories have been proposed to answer this question. One of these is 'cognitive', where primary stress is placed upon the deduction of mental processes which are thought to underlie the creation and understanding of utterances.

The Sociological View. Neither language acquisition nor language behaviour occurs in a social vacuum. In fact, both are tied to the social roles and situations of the speaker-listener. Much study on this question has been focussed on those aspects of language which vary in different social situations.

Communications
Living is largely a matter of communicating. The husband kisses his wife, the customer looks at the price tag; the pupil raises his hand; the little girl smiles. They are all communicating. People communicate from morning to night, particularly in the modern world, where most people make their living communicating. Even when communication is not the principal line of work it is still a principal part of peoples' lives. We all communicate through codes of human interaction and through language. As speech involves production of sound waves it cannot be conveyed in an acoustical mode over moderate distances without disturbing others and losing privacy. Over large distances the human voice becomes inadequate while acoustic amplification of speech will generally be
unacceptable in modern society. As a result, to communicate over long distances we need to resort to electrical techniques. In this case acoustical to electrical and electrical to acoustical transducers are used. The former transforms the speech into an electrical format whilst the latter is used at the distance point to reconvert the electric signal back into its acoustic form. One such means of long distance speech communication in this manner is by the telephone.

The telephone was invented by the Scottish-born Alexander Graham Bell in the USA in 1876, and was demonstrated at a meeting of the British Association in Glasgow the same year. At first only point to point connections were possible between stations not more than a few miles apart. The first telephone sets used Bell receivers as microphones as well as using them as earphones and so the transmission performance was rather poor with an extremely narrow and peaky frequency response. Telephone exchanges were then opened to interconnect customers and before the end of 1879, three exchanges were being used to serve about 200 customers by overhead wires.

Today, the telephone system is still a rapidly expanding network. The increasing variety of plant and equipment in telephone networks has now necessitated a gradual change in the criterion for planning from one based purely on ensuring sufficiently small losses of
loudness of received speech to the modern assessment criteria which must take into account the wide variety of transmission factors now requiring to be controlled.

The present stage in this development process is not final; already the use of extensive digital transmission networks being realised has shown the need for much more attention to be given to factors such as attenuation/frequency distortion which, in the past lossy networks, were of only minor importance.

If one compares the quality of speech obtained via the telephone system and that conveyed for example via the VHF broadcasting radio network it would be generally found that the latter is far superior. Why is this so? Well, as we have indicated, the telephone system imposes far more degradation on the transmitted speech (300-3420Hz) than does the radio broadcast network. Apart from the non-linear distortion of the speech amplitude and the departure from the flat frequency response, a telephone link applies a greater bandwidth restriction on the speech than does the radio network. The effect of this is to reduce the received speech quality by suppressing the low frequency base sounds and the high frequency sibilant sounds which are normally enjoyed when listening to a VHF radio receiver. The imposition of the transmission channel bandwidth limitations by the telephone system (or any other speech link) is of primary concern to our work and we will present some notions in this thesis which may be used to overcome the problem of speech quality reduction due to bandlimited transmission media.
1.2 Organisation of Thesis

A brief scenario of each of the following chapters in this thesis is presented.

Chapter II is a 'foundation' chapter; it includes the topic of speech communications and also deals with the subject of the mechanisms of speech production and perception in fairly broad detail. After outlining the different types of speech waveform produced we will then study the field of voice communications with a specific bias towards a description of the civil telephone system and re-affirm the various forms of degradation imposed upon the speech signal by such a network. After this we will then describe two specific applications at which the results of this project may be directed, namely, audio-teleconferencing and broadcast 'phone-in' programmes.

Since the topic discussed in this chapter may be fairly familiar to some readers, such readers may omit this chapter without losing the continuity of the thesis.

Chapter III is a review chapter of processing techniques applied to speech signals. The reason for including this chapter is two-fold:

(i) To acquaint the non-specialised reader with the existing speech processing and coding techniques and to compare them.

(ii) To provide the necessary background knowledge and to establish the framework for the investigations which follow.
The survey begins with a brief presentation of waveform coding techniques. The different quantization strategies are discussed in Pulse Code Modulators and the differential coding methods are explained including some forms of delta modulation. We then discuss an alternative approach to waveform coding which relies upon basic modelling of the speech process. The methods used under this heading are termed 'Analysis - Synthesis' Techniques and these are applied to narrow band speech signals. Afterwards, the two categories are combined to form the intermediate coding schemes i.e. those which rely upon speech modelling and waveform encoding. We then see how these types of processes can be extended to the application of wideband speed signals to enable bandwidth compression and extension to be performed. Finally in this chapter, we briefly review some examples of voiced/unvoiced detection procedures applied to telephone bandwidth and wide band speech signals.

Chapter IV describes some of the techniques we have used in order to define whether the speech excitation is voiced or unvoiced. Before explaining some of the details behind these techniques we illustrate the procedure adopted to perform the experiments upon which the results of this thesis are based. Since the experimental details do not warrant a complete chapter, we have included them as a sub-section to this chapter. We will then move on to describe some of the voiced-unvoiced classification strategies such as signal waveform inspection; spectral distribution; detection based upon power level measurements and finally we invoke the first shift
autocorrelation coefficient to classify voiced speech. The methods discussed are by no means exhaustive but they are generally used as the initial stage from which the bandwidth compression and extension schemes considered later proceed. The results and parameter settings obtained from our detection schemes are entirely based on our own speech data and are not necessarily considered applicable to the general case. Far more work and testing obviously would be required to meet this latter objective.

The thesis is then effectively divided into two main areas, namely speech quality enhancement by bandwidth extension techniques applied at the receiver and speech quality improvement by bandwidth compression/expansion techniques applied at the transmitter as well as the receiver. These two categories are conveniently described in Chapters V and VI respectively.

In Chapter V, we start by applying a spectral duplication process whereby the received telephone bandwidth spectrum is simply translated to a higher frequency position, scaled and then added back to the inband spectrum. We then present a more signal specific arrangement where the high frequency excitation signal is still derived from the received signal but is then spectrally shaped into a high frequency 'formant' before being added to the voiced speech. Our attention is then directed to the application of a similar process to unvoiced sounds using different spectral shaping. In doing this we have looked at different methods of generating the high frequency excitation signal and using them for various unvoiced
signals. Since these latter schemes require side-information regarding the duration of the unvoiced sounds, we have investigated some digital techniques to encode this side information in order to multiplex it with the bandlimited speech signal. Some computer simulated results which compare these digital codecs with and without bandwidth enhancement post processing is presented.

Finally in this chapter, we apply low frequency regeneration of the spectral components below 300Hz and add this to the 300 to 3400Hz bandlimited speech signal.

Chapter VI concentrates on the second of our quality enhancement categories where essentially bandwidth compression is applied to the wideband input speech in order to transmit a signal occupying the telephone bandwidth. At the receiver, the inverse process is applied in order to attempt to reproduce the original wideband signal. Most of our processing techniques apply no bandwidth compression or expansion to voiced speech, i.e. they merely allow the voiced speech to be bandlimited by the transmission channel. Our emphasis is placed upon the compression of the wideband unvoiced speech component whose significance to the quality of wideband speech is considered higher than that of voiced speech as far as the subjective impression of bandwidth is concerned.

Our first technique in this area is termed the voiced/unvoiced band switching system in which the wideband speech is divided into two bands, the band having the largest energy being transmitted to the
receiver. We then consider processes applied in the frequency domain where an explanation of the consequences of using window weighting functions applied to block segmented data is presented. A bandwidth compression technique using the second order spectrum (taking two successive discrete Fourier transforms of the same block of data) is then described. Another compression technique that we term 'frequency mapping' is presented in which only the most significant spectral components are transmitted with full integrity, and the remaining components are decimated in frequency prior to transmission. Initially, the spectral regions upon which the frequency components are selected are fixed, but the algorithm is then made more signal specific by adjusting these spectral regions and compression characteristics to suit the spectrum of the input signal, this we term adaptive frequency mapping. We subsequently perform another spectral compression scheme whereby the magnitude spectrum is left unaltered but the phase spectrum is now processed. We then return our attention to processing the signal in the time domain by adopting a scheme known as 'time domain harmonic scaling' proposed by Malah. (ref 2) Our objective in this case is to process wideband unvoiced signals as opposed to telephone bandwidth voiced speech.

The bandwidth compression schemes, whether performed in the time or the frequency domain, all transmit a time domain signal which occupies the telephone bandwidth. We are thus able to exploit again some of the currently available digital coding techniques in order to transmit this signal and compare the coding strategies at different transmission bit rates with and without the bandwidth compression and expansion systems.
Finally in Chapter VII, the main results of the systems reported in this thesis are compared using segmented SNR measurements (determined both in the time and in the frequency domains). The systems are also evaluated by informal subjective listening experiences and spectrographic displays of the computer simulated processed signals. Some recommendations are also indicated regarding future work.

The overall arrangement of the thesis is illustrated in figure 1.1.

1.3 Summary of Main Results

The main results presented in this thesis are outlined as follows: First in Chapter IV we perform some simple examples of voice-unvoiced switching procedures to be used in conjunction with the systems described in the following two chapters. The parameter settings decision strategies have been found to operate fairly well on a ten word sequence of speech used to test the bandwidth compression and expansion schemes.

In Chapter V we concentrate on improving the quality of bandlimited speech by spectrally extending the high frequencies of the signal at the receiver. The spectral duplication method is seen to produce a signal that subjectively appears to have a wider bandwidth than the received signal but is very distorted and of very poor quality. When high frequency regeneration is applied in a more signal specific manner, by adding an extra formant to the received bandlimited voiced speech, the resultant signal subjectively sounds
Acknowledgements
List of Symbols
Table of Contents
Synopsis

Chapter I
Organisation of Thesis

Chapter II
Speech Communication

Chapter III
Speech Processing Techniques

Chapter IV
Voiced/Unvoiced Excitation Classification

Chapter V
Speech Bandwidth Enhancement

Chapter VI
Bandwidth Compression of Wideband Speech

Chapter VII
Recapitulation

References
Appendix

Figure 1.1 Thesis Layout
more like a spurious 'whistle' than a valid component of the speech signal. When our attention is directed to bandwidth enhancement of unvoiced sounds, high frequency regeneration of the excitation signal may be achieved by applying various non-linear transfer functions to the received in band speech and then spectrally shaping the upper frequencies before adding them back to the received speech. By allowing this process to operate on particular unvoiced signals from our data sequence some noticeable improvement was observed giving the subjective impression of a wider bandwidth than that actually used for transmitting the bandlimited speech. There was also only a minimal amount of distortion observed in the processed speech signal. By replacing the bandlimited channel with a digital medium, better quality speech was also obtained at a given transmission bit rate with the aforementioned system rather than without using any post-processing at all.

The results obtained for the regeneration of the spectral components below 300Hz showed no overall improvement in quality compared to the subjective impression of the 300 to 3400kHz bandlimited speech although there was a perceptual presence of the base frequencies which were absent from the bandlimited signal.

In Chapter VI we find that invoking the transmitter terminal for processing the wideband input speech signal tends to offer more potential for producing better subjective quality of speech at the output of the receiver terminal than the systems discussed under the framework of Chapter V. The introduction of the voiced-unvoiced band switching system (VUBS) had shown a subjective impression of a
broader bandwidth of the output signal than a 300 to 3400Hz signal that would normally be conveyed by the transmission channel. There was however some unnaturalness associated with the switching from voiced to unvoiced speech and vice versa. The elaboration of the pre- and post-processing system in order to render this switching less obtrusive shows that it does seem possible to produce a more natural sounding output signal. This was achieved by transmitting more information about the input signal spectrum than that provided by the VUBS system. The schemes which show improvement in regard to this latter strategy are the random phase processor, the time domain harmonic scaling system (ref 2) and the frequency mapping procedure; this was particularly notable for the adaptive version of the frequency mapping process.

Once again, we find that the replacement of the 300 to 3400Hz analogue channel by a digital link seems to provide a better quality speech signal, when the spectral compression systems are used, than the speech signal obtained without any pre- and post-processing.

In Chapter VII where some of the systems are compared on the basis of SNR measurements and informal subjective listening tests, we find that generally the bandwidth compression/expansion schemes of Chapter VI tend to give an improved performance over the bandwidth extension techniques described in Chapter V.
2.1 Speech Production and Perception (ref 3)

2.1.1 Speech Production

Speech is the acoustic end-product of voluntarily formalised motions of the respiratory and masticatory apparatus. The process must be learned, developed, controlled and maintained by acoustic and kinaesthetic feedback of the speech musculature. Information from these senses is organised and co-ordinated by the central nervous system which is used to direct the speech function. Impairment of either control mechanism usually degrades the performance of the vocal apparatus - this is apparent with partially or totally deaf people. It has been thought that speech evolved when ancient man discovered that he could supplement his communicative hand-signals with gestures of his vocal tract.

Figure 2.1 represents a mid-Sagittal section through the vocal tract of an adult. The primary function of inhalation occurs by expanding the rib cage, reducing the air pressure in the lungs which draws air into the lungs via the nostrils, nasal cavity, velum part and windpipe (trachea). Air is normally expelled by the same route. In eating, mastication takes place in the oral cavity. The vocal tract is an acoustical tube, variable in cross sectional area which is non-uniform. The tract is terminated by the lips at one end and the vocal chords at the other. It is deformed by movement of the
articulators, lips, jaws, tongue and velum. The vocal tract is about 17cm long for the adult male and the cross-sectional area varies from 0-20cm². The nasal cavity constitutes an auxiliary path for sound transmission. The velum-to-nostrils length is about 12cm and the volume is about 60cm³, and it is also partly partitioned by the nasal septum. The acoustic coupling is controlled by the size of the opening at the velum. This coupling between the nasal cavity and mouth can substantially influence the character of sound radiated from the mouth; with non-nasal sounds, the velum is drawn tightly up.

The lung pressure varies between 4-20mm H₂O during speech which is maintained by a steady slow contraction of the rib cage. The slit-like orifice between the vocal chords is called the glottis.

Varied sounds of speech are produced by the vibratory action of the vocal chords which is called phonation. The tense chords are initially together whereby the subglottal pressure is increased sufficiently to force them apart with lateral acceleration. The airflow builds up in the orifice and the local pressure is reduced according to the Bernoulli relation and the elastic force then acts to return the chords to a proximate position. The chords are again drawn together, the flow diminishes and the local pressure approaches a subglottal pressure whereby the relaxation cycle is repeated. The mass compliance and subglottal pressure essentially determine the period of oscillation which is shorter than the natural period, hence the chords are thus driven in forced
oscillation. The air flow through the glottis as a function of time is similar to the area or the glottal opening function versus time function. The normal glottal waveform is roughly triangular in shape and therefore has a frequency spectrum rich in harmonics diminishing in amplitude at about 12dB per octave. The maximum glottal area, however, is correlated with voice intensity to a surprisingly small extent. Because of the small glottal opening, the acoustic impedance of the glottal source is high compared to that of the vocal tract.

Turbulent flow of air created by some point of structure in the tract causes another source of vocal excitation. This is incoherent excitation and as the associated spectrum is uniform at the region of generation, it results in a continuant sound. The vocal cavities in front of the constriction are usually the most influential in spectrally shaping the sound.

A third source of excitation is created by a pressure build-up at some point of closure. An abrupt release provides a transient (approximating to a step function) excitation to the vocal tract. This might be considered to have a spectrum proportional to \(1/f\) (transform of a step function). In this way unvoiced plosive sounds are produced. Whispered speech is produced by substituting a noise source for the normally vibrating vocal chords. The source may be produced by a turbulent flow at the partially closed glottis or some other constricted place in the tract.
The Sounds of Speech

To be a practicable medium for the transmission of information, a language must consist of a finite member of indistinguishable mutually exclusive sounds. When one unit replaces another in an utterance, the meaning changes. The basic linguistic element is called a phoneme which might be looked upon as a code uniquely related to the articulatory gestures of a given language.

In continuous articulation, the vocal tract dwells only momentarily in a state appropriate to a given phoneme. Despite the mobility of the vocal apparatus and continuous nature of the speech wave, one can subjectively segment speech into phonemes. Phonetic transcriptions are usually enclosed in slashes //.

The classification of speech sounds is made according to the manner and place of production. The articulation of vowel sounds is generally described by the position of the tongue hump along the vocal tract and the degree of constriction. Vowels are normally produced exclusively by voiced excitation of the tract; the tract being maintained in a stable configuration during most of the sound. Vowels are further characterised by negligible nasal coupling, radiation being only from the mouth otherwise the vowel is nasalised. The 12 vowels are shown in Table 2.1; the approximate articulatory configurations are shown in figure 2.2. The shapes of the pharynx cavity and lower vocal tract may be deduced from X-rays.
Consonants constitute those sounds which are not exclusively voiced and are mouth radiated from a relatively stable configuration characterised by greater tract constrictions than the vowels. The sustained sounds are called continuants.

Fricative consonants are produced from an incoherent noise excitation in the vocal tract. Noise is generated by a turbulent air flow at some point of constriction e.g. tongue behind teeth (dental); the upper teeth on lower lip (labio-dental); the tongue on gum ridge (alveolar), the tongue against hard or soft palate (palatal or velar) and the vocal chords fixed (glottal). Radiation of fricatives normally occurs from the mouth. They can be voiced or unvoiced. Given an articulatory configuration, the voiced and unvoiced fricatives are paired as cognates as shown in Table 2.2.

Stop consonants depend upon vocal tract dynamics for their creation. The lungs build up pressure behind a complete closure in the vocal tract (occlusion), the pressure is then released by an abrupt motion of the articulators. The explosion and aspiration of air helps to characterise stops. Again, stops have cognate pairs shown in Table 2.3 and figure 2.3, each position is just prior to pressure release.

Nasal consonants are normally voiced with a complete closure being made towards the front of the vocal tract (lips, tongue on gum ridge or tongue on hard palate). The velum is open wide and the nasal tract provides most of the transmission channel with radiation at
the nostrils. The closed oral cavity acts as a side branch resonator coupled to the main path which can substantially influence the sound radiated. Because the nasal consonants are sustained, they are classed as continuants; these are listed in Table 2.4 and their vocal profiles are illustrated in figure 2.4.

Two small groups of consonants greatly resemble vowels, these are glides /w,j/ and semi-vowels /r,l/. They are characterised by voiced excitation of the vocal tract with no effective nasal coupling and mainly radiation from the mouth. Glides are dynamic sounds which invariably precede a vowel and exhibit movement towards a vowel. The semi-vowels are continuants in which the oral channel is more constricted than in most vowels and the tongue tip is now up. These sounds are listed in Table 2.5 and their profiles for the beginning positions are given in figure 2.5.

Diphthongs and affricatives are combination sounds where some of the preceding vowel or consonant elements combine to form basic sounds whose phonetic values depend upon vocal tract motion. Pairs of vowels appropriately combine to form diphthongs which are vowel-like in nature. They are characterised by a change from one vocal position to another e.g. from /e/ to /I/ forms /ei/ as in 'say' and /IU/ as in 'new', /aI/ as in 'boy', /au/ as in 'out', /aI/ as in 'I' and /OU/ as in 'go'.

Stop-fricatives form affricates when combined such as /tʃ/ as in 'chew' and /dʒ/ as in 'jar'.
2.1.2 The Ear and Hearing

The ultimate recipient of information in a speech communication link is usually man. His perceptual abilities dictate the precision with which speech data must be processed and transmitted. These abilities essentially prescribe fidelity criteria for reception and in effect determine the channel capacity necessary for the transmission of voice messages.

The acoustic-mechanical operation of the peripheral ear has been put on a firm basis by G. V. Bekesy (Ref 4). However, in contrast, present knowledge is relatively incomplete about the inner ear. The function of the ear is to convert mechanical motion into neural activity. The primary acoustic transducer of the human is shown in figure 2.6 where it is divided into three regions; the outer ear, the middle ear and the inner ear.

a) The Outer Ear

The term 'ear' usually applies to the convoluted appendage on the side of the head. This structure is the pinna and it surrounds the entrance to the ear canal. Its main function is to protect the external canal and its directional characteristics at audio high frequencies probably facilitate localisation of sound sources.

In man, the external canal (meatus) is about 2.7cm in length and 0.7cm in diameter being terminated by a thin membrane which is the ear drum (tympanic membrane). This has the form of a relatively
stiff inwardly directed cone with an included angle of about 135°. To a rough approximation, the meatus is a uniform tube, open at one end and closed at the other. The first normal mode of vibration occurs at about 3000Hz.

b) The Middle Ear

Just interior to the ear drum is the air filled cavity which contains the ossicular bones to provide impedance transformation i.e. to convert the external sound pressure into a fluid volume displacement in the inner ear. The hammer (malleus) is fixed to and rests on the ear drum. It makes contact with the anvil (incus) which in turn connects via a small joint to the stirrup (stapes). The footplate of the stirrup seats in a port, the oval window, and is retained there by a annular ligament. The oval window is the entrance to the inner ear.

The middle ear structure also provides protection against loud sounds which may damage the more delicate inner ear.

c) The Inner Ear

The inner ear is composed of the cochlea (figure 2.7) which is normally coiled into 2 1/2 turns. This is where mechanical to neural transduction takes place. The cochlea is filled with a colourless liquid having a viscosity about twice that of water and a relative density slightly greater than unity.
The cochlea is divided along almost the whole of its length by a partition. The partition itself is a channel bounded by a gelatinous membrane called the basilar membrane and another membrane called REISSNER'S membrane.

The inner ear is connected to the middle ear by the stapes footplate supported by an oval window. In vibrating, the stapes act as a piston producing a volume displacement of the cochlea's fluid. This fluid is compressible and the cochlea is essentially rigid. The fluid displacement is caused by inward displacement of the stapes which must be relieved and this is accomplished at the round window which is covered by a compliant membrane. Very slow vibration of the stapes (< 20Hz) result in to-and-fro flow of the fluid through the opening at the helicotrema (figure 2.7). Higher frequency vibrations are transmitted through the yielding cochlea partition at a point which depends upon the frequency content of the sound stimulation.

A cross section of the cochlea and its partition is shown in figure 2.8. The organ of corti rests on the basilar membrane which contains some 30,000 sensory hair cells on which the auditory nerve terminate. The basilar membrane is stiffer and less massive at its narrow end and more compliant and massive at its broader end. Its resonant properties therefore vary continuously along its length.

Von Bekesy vibrated the stapes footplate sinusoidally and measured the amplitude and phase of the membrane displacements along the length of the cochlear. The amplitude and phase response of a given
membrane point is much like that of a relatively broad-band filter. The amplitude responses of successive points are roughly constant - Q in nature. Because of the constant percentage bandwidth property, the frequency resolution is best at the low frequency end of the membrane and the time resolution is best at the high frequency end. Linear increments of distance along the basilar membrane correspond to approximately logarithmic increments of peaks of the frequency response for resonant frequencies up to about 1000Hz. Excitation of the stapes is propagated down the membrane in the form of a travelling wave of displacement. Because of the taper of the distributed constants with distance, essentially no reflection takes place at the helicotrema and no standing wave of displacement is created (longitudinal). The membrane is a dispersive transmission medium, and so the travelling wave is damped in hf components as it progresses towards the helicotrema and its group delay increases.

Mechanical to neural transduction takes place in the organ of corti (containing the hair cells). The hair cells are in contact with a third membrane - the tectorial membrane. A deformation of the basilar membrane causes relative motions between the tectorial membrane and the reticular limina (through which the hairs protrude) causing a resultant stress on the hairs between them. By a process that is presently not understood, a bending of the hairs produces an electrical discharge in the cochlear portion of the VIIth Caranial nerve. (ref 5) The sensory cells are connected to the brain via the bundle of nerve cells (neurons) comprising the auditory nerve containing more than 30,000 neurons. Neurons presumably have two
states, namely active and inactive. When excited by an electrical input above a particular threshold, they produce a standard electrical pulse of about 1mS duration and are desensitised for a period of 1-3mS thereafter. They can be consequently excited to a maximum discharge rate of the order of 300 to 1000S⁻¹. Excitation of the neurons in accordance with a time waveform of a frequency greater than 1kHz is thus not possible. The cochlea is consequently considered as an intensity spectral analyser.

The fact that neurons leading away from a given region in the frequency selective basilar membrane and maintain their identity within the auditory nerve offers a possibility for the coding of pitch in terms of membrane place of maximum intensity.

2.1.3 Loudness (ref 6)

Loudness is defined as the magnitude of the auditory sensation which a sound produces. The threshold of hearing intensity varies over the frequency range which shows that the loudness of a sound does not solely depend on its intensity. Generally, a tone of around 1000Hz produces maximum auditory sensitivity.

It is therefore evident that as loudness depends on frequency as well as intensity, the decibel cannot be used as a loudness measurement unless it is modified in some way since the decibel is a ratio of two intensities. In assessing loudness, subjective techniques must be used with the listener judging the effect of
changing intensity or frequency. This leads to a unit of loudness level called the 'phon'. In calibrating loudness by the phon the technique is to compare the loudness of a reference tone with the loudness of the given tone. The intensity of the reference tone can be adjusted until it sounds as loud as the given note. The amount by which the reference tone has moved away from its threshold of hearing intensity value gives the loudness level of the given tone in phons. At 1000Hz where there is 120dB between the threshold of hearing and the threshold of pain, there are 120 phons; the phons and the decibel are taken to be equal (at the reference frequency).

Referring back to the auditory process, the greater the sound intensity, the greater the number of auditory neurons activated. Relative levels of typical sounds (in dB) are exhibited in Table 2.6.

2.2 Characterisation of Speech Waveforms (ref 7)

After the speech sounds have been produced, their waveforms can be observed by recording the speech sound via a microphone and playing it back to an oscilloscope. Unless a storage oscilloscope is used, it will be difficult to distinguish any part of the waveform since the irregularity of the speech wave proves too difficult for the synchronisation of the time base trigger circuit. If, however, the speech signal is lowpass filtered and time sampled at a finite rate (the Nyquist rate for a lowpass signal is twice the highest frequency therein), the time samples may then be amplitude quantised and stored in digital form. The digital samples can be plotted to
form a static speech waveform as shown in figure 2.9. The waveform oscillogram immediately puts into evidence the differences between the relatively intense, quasi-periodic voiced sounds of speech (produced by the periodic vibration of the vocal chords) and the lower amplitude unvoiced sounds (produced by the random noise of turbulent air flow through a constriction). The unvoiced sound appears much more random in structure than the voiced waveform.

The average probabilities of different frequency components in speech are illustrated by the long-time-averaged spectral density i.e. the \( S(e^{j\omega}) \) plot of figure 2.10 indicating the Power Spectral Density. It is clear that high frequency waveform components contribute very little to the total speech energy. Nevertheless, these high frequency components are very important carriers of speech information and as such they must be adequately represented in coding systems. Unvoiced waveform segments have highpass spectra whilst voiced segments, although they globally exhibit lowpass frequencies, display local resonances from the vocal tract called formants. In fact, the temporal variation of the formant frequencies (formant trajectories) is extremely significant to speech signal intelligibility. This shows that short-time speech spectra do not always permit simple lowpass descriptions as with the long-time averaged spectral density of speech. Figure 2.11 shows the short-time spectra for typical voiced and unvoiced segments.

The short-time spectrum due to the voiced signal exhibits spectral envelope peaks due to the resonant formants as well as a 'fine-structure' which is regular due to the harmonics of the
fundamental pitch period. The fine structure is indicative of the 'excitation' of the utterance whereas the spectral envelope displays shaping by the vocal-tract response.

The short-time spectrum for the unvoiced sounds, particularly the /s/ sound, exhibits a random-like fine structure due to the turbulent excitation. The spectral envelope has a peak, typically between 4 and 5 kHz, due to the resonance of the tract between the place of constriction i.e. the tongue and upper teeth (which is the source of the excitation) and the opening at the lips.

Characteristics of specific unvoiced segments will be dealt with in more detail in section 4.3.

2.3 Voice Communications

Communication is the exchange of information. Sometimes, however, information must be conveyed over long distances almost at once if it is to be of any use; a distress signal from a sinking ship or a telephone call to arrange a meeting with a friend are two examples of this. For these we need the special forms of communication known as telecommunication. The ancient Greek word 'tele' means 'far', so the word 'telecommunication' means literally 'communication over long distances.

There are, of course, many ways of communicating over long distances, e.g. smoke signals, beacons, drums and semaphore. For engineering purposes, we usually take telecommunication to include
only the electrical systems of long distance communication. Smoke signal technology has little in common with the telephone! Telecommunications can therefore be defined more fully as sending information by electrical means over distances which can be greater than the normal range of the senses.

In this section, we are concerned specifically with telephony (from the Greek word 'phone' meaning sound) which enables speech to be conveyed beyond the range of the human voice and hearing. In telephony, messages are sent and received in spoken form using instruments called telephones. The sound waves comprising speech are used to control the transmission of electrical signals in the form of alternating currents at the same frequencies as the sound waves and these are used to reproduce the original sounds at the receiving end. The telephone is easy to use without special training which means that people making calls can converse directly with each other but there is not normally any permanent record of messages sent or received to verify a message perhaps later. Both parties must generally be available for the call at the same time.

2.3.1 The Telephone System (refs 8,9)

The majority of telephone users generally do not have a great deal of knowledge or consideration of the functioning of the telephone system; in most cases it is usually taken for granted except of course, when it is blamed for getting a wrong number, providing a 'bad line' or possibly the most infuriating, getting 'cut off'.
This attitude is also generally the same as with other service industries such as electricity supply, gas, water and sewage i.e. they are generally taken for granted unless 'something goes wrong'.

The telephone system has existed in the UK for nearly 100 years. During this time it has developed into a system which enables a subscriber to converse with another either in his own town, or in a matter of seconds, and without any intervention of any human operator, to another subscriber in a distant part of the world and converse with him as easily as if they were talking face-to-face. On an international scale, uniformity of standards is maintained, as far as possible, by the Comite Consultatif International Telegraphique et Telephonique (CCITT).

A system such as the British telephone system might contain, typically, 20 million terminals, 21 million transmission links and 6600 exchanges. In the trunk network, although it is known what the nature of the connection between any two points is, it is not known what its precise characteristics are; when a call is connected by automatic subscriber trunk dialling (STD) there is no way of knowing by what route the connection will be established as this depends upon the circuits available at the time. So, even if the characteristics of every link in the network were known, it would not be possible to predict what would be the characteristics of the connection between any two given subscribers.

A telephone connection starts with the talking subscriber who is provided with a telephone set connected to the local exchange by a
twisted pair of wires. The connections between local exchanges are known as junction circuits; longer distance connections may be made over the trunk or transit circuits. Because of the distances involved, the trunk circuit will often require amplification as it becomes economic to multiplex many speech channels on to a single cable or radio link rather than using twisted pairs for each connection.

2.3.1.1 The Telephone Set

All modern subscribers' telephone sets are handset telephones in which the microphone and earphone are mounted on either side of a handle connected by a light flexible cable to a base unit containing all the other components and upon which the handset rests when not in use. The microphone used in all telephone sets from the early days of telephony until the last 2-3 years has been the carbon granule type. The carbon granule microphone is an imperfect transducer of speech sounds, but it has an advantage over nearly all other types of microphone in that it acts as an amplifier; the resistance across the granule chamber is modulated by the incident sound wave and this in turn modulates the feeding current passing through it, producing a signal with energy several orders of magnitude greater than that which was incident upon the microphone diaphragm.

The earphone used in most modern telephone sets are of the 'rocking-armature' type. The armature vibrates in response to the variations of the magnetic field set up by an electromagnet.
All the other components are in the base of the telephone set, some of which comprise the anti-sidetone circuit. If the microphone and earphone are connected in a simple way to the two wires of the subscribers' line, the talker will hear his own speech considerably louder than that of the distant subscriber's speech. This makes it difficult for him to hear the distant subscriber so he tends to lower his own voice to overcome this. The anti-sidetone circuit tends to reduce the amount of the microphone signal which appears in the earphone circuit. Another important function of the speech circuit is to regulate the transmitting and receiving efficiencies depending upon the length of the subscriber's line.

The complete circuit of the Post Office telephone No 706 is shown in figure 2.12(a). It shows the switching and signalling components as well as the speech circuit. As the telephone set must be designed for absolute minimum costs, the two are not entirely independent. The diagram shows the telephone set in the 'on-hook' condition i.e. $S_1$ and $S_2$ contacts are open. Only the bell is connected across the line in series with a dc blocking capacitor. Ringing current for the bell is an alternating current at $50/3$ Hz and about 75 volts rms.

When the handset is lifted, $S_1$ and $S_2$ close and a dc path (subscriber's loop) is connected to line. This subscriber's loop has the effect of signalling to the exchange either a call request or ringing-trip depending upon whether the call is to be originated or received. When the dial is rotated from its normal rest
position, the switch $S_3$ is closed so that the dial pulse springs $S_5$ are applied directly across the line. $S_5$ is also closed to short circuit the earphone receiver to prevent dial clicks being heard. When the dial is released, a toothed wheel and trigger assembly under the finger plate causes an appropriate number of loop-disconnections and re-connections by opening and closing $S_5$ at 10 pulses per second. During dialling, the spark quench circuit in the telephone prevents sparking at the pulse spring contacts and the bell shunt circuit prevents bell-tinkle. The anti-sidetone circuit, figure 2.12(b), functions by setting up opposing currents to minimise the emf across its earphone when the microphone is producing an alternating current. The appropriate circuit is known as the balance impedance network. The impedance of the bell at voice frequencies is relatively high, hence it has little consequence in the balance impedance circuit.

The regulator, figure 2.12(c), consists of a resistance in parallel with a diodes chain which is connected in series with the microphone such that the microphone current passing through the resistors biases the diodes. As the line length decreases, the line current increases and the impedance of the diodes decreases. The ballast resistor is a tungsten filament lamp whose resistance is low (10 ohms) on long lines and higher (40 ohms) on short lines so that the regulation effect is enhanced.

The loud speaking telephone has characteristics which are very different from those of the ordinary handset telephone. It is provided with an amplifier and loud speaker for receiving and a linear microphone and amplifier for transmitting.
To prevent the received signal from being re-transmitted, which is not only objectional to the distant subscriber but may cause instability (howl-sound), voice-operated gain adjusting has to be included in both transmit and receive paths. The speech signal is affected by this, also the speaker is talking at some distance from the microphone in a room which may be highly reverberant.

2.3.1.2 The Subscriber's Line and Local Exchange

Each subscriber is connected to the exchange equipment by a twisted pair of wires. The cables linking the exchange in densely populated urban areas may each contain hundreds of twisted pairs and precautions are necessary to prevent mutual interference of signals between pairs. For example, neighbouring pairs are twisted with different pitches. Sometimes two subscribers share a single pair (shared-service).

Once a call has been established, the only function of the exchange equipment is to supply power to the subscriber's telephone set, this being done using the Stone bridge feed circuit, shown in figure 2.13. The line is fed from the 50 volt exchange battery with a positive earth. To minimise interference noise induced into the line, the subscriber's line is completely balanced with respect to earth by the inductor which has two equal windings and isolates the speech from the battery.
Junction circuits are lines which connect two local exchanges. Pulse code modulation (PCM) is sometimes used to carry 24 circuits on 2 twisted pairs (1 pair for each direction) where they could only carry two before.

Trunk circuits are used for calls outside the local area. The trunk network serves only to interconnect local exchanges. Because the distance involved may be very large, a wider variety of transmission methods is used in the trunk network.

The possible routes of a trunk connection are shown in figure 2.14. When amplification is used, separate go and return paths must be provided as shown in figure 2.15(a). To make this possible, signals in each direction on the two-wire trunk junction circuit must be separated which is done with a hybrid transformer (Figure 2.15(b)) which operates in a manner similar to the anti-sidetone circuit of the subscriber's telephone set. Each port of the hybrid looks into the same impedance of 600 ohms and is connected to windings of the same number of turns. There is no net voltage across the port where two equally opposing voltages are developed in the primary windings, hence, a signal entering any port divides equally between the two neighbouring ports and causes no signal in the opposite port. In order to achieve this the balance impedance must match the 2-wire line whose impedance is frequency dependent, complex and also depends upon the other lines to which it is connected.

Frequency division multiplexing is used for the majority of trunk circuits using a single sideband suppressed carrier with channel
spacing of 4kHz. The channels are modulated up into groups of twelve. The groups are again modulated into supergroups of 5. The supergroups are then modulated into Wideband Systems of 3. The smaller groups can be transmitted over twisted pair cables, but for the larger groups, coaxial cables and line-of-sight microwave systems are used. For international calls, submarine cables or synchronous orbital satellites are used employing time assignment speech interpolation (TASI).

Long circuits introduce a problem of delay. When the delay is very long it accentuates the problem of echo caused by reflections at the 2 to 4-wire transitions in a connection. These reflections cause the speaker to hear his own speech delayed. If the delay is long, he hears one syllable repeated while he is trying to utter the next which is disruptive to his speech. To alleviate this effect echo suppression, switched alternation and more recently echo cancellation (ref 10) are used in go or return paths.

2.3.2 Causes of Degradation of the Speech Signal by the Telephone System

a) The Talker

When a talker speaks on the telephone, he speaks differently than when in a face-to-face conversation which may make a difference between speech signals observed on the telephone line and those obtained under laboratory conditions. The reduced intelligibility of signals transmitted over the telephone network, particularly that
of certain consonant sounds caused by the high frequency cut-off of
the system, may cause speakers to articulate their words more
carefully than they would in face to face speech, particularly when
the speech that they hear from the distant subscriber is of reduced
intelligibility. A talker also tends to use greater vocal effort
when speaking on the telephone and this affects the quality of the
speech. The loudness of a talker's speech depends upon a number of
different social factors; talkers in densely populated areas talk
louder; long distance calls are louder than local calls regardless
of the loss in the connection, men talk louder than women, business
calls are louder than social calls.

The handset position considerably affects the loudness of the
transmitted speech. The manner of holding the microphone can vary
from holding it under the chin to holding it in front of the mouth
with the spare hand cupped round the mouthpiece. This also affects
the frequency response of the airpaths from the mouth and nose to
the microphone.

b) The Talker's Environment

The majority of telephone calls are normally made under fairly quiet
conditions. However, telephone kiosks are often sited in public
thoroughfares perhaps only a few feet from a road carrying heavy
traffic. Due to the environmental noise, the signal-to-noise ratio
at the transmitter will be very low.
With the close-talking 'microphone of the normal telephone handset, the acoustic properties (reverberation) of the talker's environment are not normally important. With a loud speaking telephone in an office containing no sound absorbing objects other than the speaker himself, the frequency structure of the transmitted signal will be altered due to standing waves set up in the room and the reverberation can be long enough to modify the time structure of the signal significantly.

c) The Transducers

Both the microphone and the earphone in the telephone set introduce amplitude/frequency distortion. The modern rocking armature earphone introduces negligible non-linear distortion and does not change its characteristics with time. The carbon microphone on the other hand is non-linear, and all its characteristics - amplitude/frequency response, non-linearity and sensitivity - vary with its feed current and physical position and they change according to its immediate past history of acoustic excitation and mechanical movement.

2.3.2.1 Noise Introduced by the Telephone Network

Firstly, loss (attenuation) is the most significant factor in the planning of telephone networks as too much overall loss in a connection will make it unusable, where as too little loss in any of the component circuits of a connection can lead to instability.
Continuous random noise is inherent in the physical processes by which electronic devices operate; the causes of continuous random noise are thermal noise, shot noise and low frequency noise. Thermal noise is caused by movement of electrons due to the thermal energy of the molecules. It is dominated by other sources of noise in line systems. Shot noise is due to the discrete nature of an electric current (i.e. electrons).

Low frequency (1/f) noise is associated with the microscopic structure of the material and caused by fluctuations in conductivity. Surveys have shown that even on the worst connections a signal to noise ratio of some 30dB can be expected.

Impulsive noises occur on line transmission systems due to atmospheric discharges and in mechanical telephone exchanges due to switching currents in neighbouring equipment. Speech is, however, remarkably resistant to impulsive noise even when it occurs in fairly long bursts. The characteristics of speech and impulsive noise are sufficiently different for it to be readily ignored by the listener. Impulsive noise occurs mostly in the step-by-step (strowger) exchanges.

a) Crosstalk

This is a term used to describe unwanted signals entering a speech channel from other channels. In voice-frequency circuits any coupling between one circuit and another (including crossed-lines) will give rise to intelligible speech crosstalk. This is
unacceptable because the telephone system is expected to provide privacy of communication. Speech signals do not become unintelligible until they are about 60dB below the normal speech level in the telephone system.

When the coupling between two speech channels is highly non-linear, it gives rise to speech crosstalk which is not intelligible, but because it consists of noises having a speech-like rhythm it is more disturbing to the listener than noises with a non-speech like structure. Non-linear coupling is mainly caused by intermodulation in FDM systems.

Multi-frequency signalling tones and mains hum may also be induced into the lines. However, the threshold of audibility of the human being is about 40dB higher at 50Hz than at 1000Hz, hence, large amounts of mains hum pass unnoticed by the listener.

Quantization noise is introduced by pulse code modulation (PCM) transmission. The continuous range of amplitude levels of the speech signal must be represented by the number of code groups available. The resultant errors are small enough to appear to be a random noise superimposed upon the signal. Logarithmic coding reduces this quantization noise at low amplitude levels. For further discussion related to PCM, see section 3.2.

Any other 'funny' noise that one sometimes hears on telephone connections are probably due to fault conditions and hence should be eliminated.
2.3.2.2 Amplitude/Frequency and Group Delay Characteristics

a) The Microphone

The carbon microphone is by far the most serious cause of degradation to the speech signal in the telephone system, but until the last few years its replacement by any other sort of microphone has been out of the question. The signals from other microphones would require amplification and the cost of including an amplifier in the telephone set would hitherto have been prohibitive. However, the conversational behaviour of the telephone user has developed to allow for its deficiencies in such a way that the degradation is frequently not noticed and a high quality signal coming out of a telephone earpiece sounds positively peculiar.

The optimum transmitting characteristic, on which the telephone microphone is designed, falls at about 9dB per octave below 800Hz and is nearly flat above. The actual response of the telephone microphone falls some way short of this ideal. Its mechanical structure normally is a response made up of three resonances (well damped in modern microphones) rather than a flat response above 800Hz. Below 800Hz the response falls off, but at about 400Hz it levels out again with a loss of about 20dB which falls off only slowly before that. Above 3kHz the response falls off sharply and stays low. The holes in the microphone front plate and sometimes those of the telephone mouthpiece, are used to provide the high frequency cut-off. Insofar as the carbon microphone has a low-pass
filtering effect on the signal it could be said to degrade the speech signal, but this effect is not as important compared with other sources of frequency limitation in the transmission part of the telephone network.

When a carbon microphone is excited with a sinusoidal signal and the resistance is monitored by observing the output voltage resulting from a constant feed current, it is immediately observed that the peaks of the waveform are distorted by being smoothly compressed. The expansive peaks become sharply clipped and it appears that there is a sound pressure level above which the resistance of the carbon chamber cannot be increased no matter how large the applied acoustic signal.

As the peaks in the speech waveform are those of the formant resonances, the effect of this instantaneous non-linear distortion is to produce false harmonics and intermodulation products of the formants.

Another consistent feature of carbon microphone signals is the addition of a level of broadband background noise whose amplitude varies with that of the input signal.

b) Transmission

The inductive loading of cables reduces their attenuation throughout the voice band, but introduces loss more rapidly above the cut-off frequency of about 3kHz. The use of PCM on junction circuits
introduces a high frequency cut-off by the use of low pass filters at the input and output which gives a flat characteristic up to the cut-off frequency of 3.4kHz and falls off very rapidly thereafter.

The frequency division multiplex system used in trunk circuits use single sideband suppressed carrier transmission and therefore, to conserve bandwidth, include a high-pass filter at 300Hz as well as the 3.4kHz low pass filter. The high pass filter removes a large amount of signal energy which contributes little to intelligibility.

c) **Mismatch**

The frequency characteristic and loss of any component part of a telephone connection depend upon the impedance terminating it. The characteristics of such a component are normally quoted for image impedance terminations, i.e. when each port of the component looks into an impedance which is the same as the impedance seen looking into that port of the component. Under these conditions, the maximum transfer of complex power (volt-amperes) both into and out of the network takes place. At the interface between the telephone set and the local line or junction, a mismatch cannot be avoided at an acceptable cost.

d) **Delay**

When there are no reflections caused by mismatch in a telephone connection, then delay is of no consequence unless it becomes so great ( \( > 600\text{ms} \)) that it causes conversation difficulty. When
significant reflections occur in the connections as well as delay, not only do the resulting echoes become subjectively unacceptable at much lower values of delay, but the amplitude/frequency response and group delay characteristic are affected. This effect causes ripples in the frequency characteristic.

e) Non-Linearity

Apart from the carbon microphone, the other source of non-linearity in the voice frequency path is the earphone. However, the non-linearity of the modern rocking armature receiver is very small.

The regulator in the 706 telephone set is inherently non-linear in its operation producing odd-harmonic distortion.

2.3.2.3 Time Varying Effects

a) Speech Operated Devices

 Voice operated switches and gain adjusting devices work on the short-term power of the speech signal and are arranged to ignore noise spikes, with the result that they tend to fail to detect the sharp onset of certain speech sounds and so mutilate the speech signal.
b) **Echo Suppressor**

This switches attenuation into one direction of transmission or the other according to which speaker is talking as judged by the relative magnitude of the signals in the two directions. Whilst only one talker is speaking, it should not switch and so should cause no mutilation, but it will tend to clip the initial speech sound of one talker interrupting the other and may cause degradation if there is a lot of background noise from one end so that the speaker who is talking is continually breaking in on the background noise.

c) **TASI (Time Assignment Speech Interpolation)**

The TASI system finds a channel for a speaker every time the talker starts a new talkspurt. If the system is underloaded, the channel may not be relinquished from one talkspurt to the next, but when a new channel has to be seized, there will be a tendency for the first speech sound to be mutilated. At times of high activity, it is possible for more talkers to be talking than there are channels available and this can lead to 'freeze-out' in which an arbitrary length from the beginning of a talkspurt may be lost until a channel becomes available.

d) **Loud Speaking Telephone**

In the loud speaking telephone, the gain that must be provided to compensate for the distance between the transducers and the
subscriber's mouth and ear would be sufficient to cause instability if voice operated switching or gain variation were not provided; there will inevitably be some mutilation of the initial sounds of talkspurts. The effect of voice switching is likely to be overshadowed by that of room reverberation. Adaptive filters have been found useful in replacing the gain switching stage (ref 10).

2.3.2.4 Random Effects

a) Fading Radio Channels

All radio channels are liable to fading. On microwave links, a certain amount of fading can be accommodated by the use of AGC (Automatic Gain Control). For deep fades, protection channels are provided and a channel in a part of the spectrum suffering a deep fade will be switched to the protection channel until it recovers.

There is a possibility therefore that the shape of the cut-off at either end of the pass band of a radio channel may vary slightly when such switching takes place. Deep fades are of short duration (of the order of 1 second).

b) Frequency Offset

FDM systems, as previously stated, use single sideband suppressed carrier modulation (SSBSC), the carriers at the receive-end being synthesised from a highly stable oscillator. These oscillators are
all synchronised to a master oscillator via a system of pilot tones. The synthesised carriers are prevented from being more than a few Hz in error. A few Hz error in a speech signal is not detectable by a listener.

c) Phase Jitter and Jumps

In addition to frequency offset (sometimes known as phase roll) the synthesis of carriers on FDM systems also introduces random fluctuations of carrier phase known as phase jitter. Although this does not affect the power spectrum of the signal, changes in the phase of the carrier alter the phase characteristic of the connection so that phase jitter causes its impulse response to change continually with time. Their probability is sufficiently low for them to be of little consequence in speech transmission.

2.3.3 Audio Teleconferencing

2.3.3.1 General

Audio teleconferencing (refs 11-13) is a service that uses the telephone network, and in some cases private circuits, to provide a link between high quality audio terminals in order to facilitate remote conferencing between the terminals. This service dispenses with the telephone handsets thereby forming a direct application to wideband speech compression/expansion or bandwidth enhancement of speech signals.
Teleconferencing may be used as a substitute for travelling and hence offset the costs otherwise involved. This can also cause considerable savings in time and can be conveniently arranged at short notice. The problems of encouraging its use are not economic or technical but mainly in educating people to become familiar with its use. As travelling costs increase, teleconferencing could become as familiar a means of communication as the ubiquitous telephone and so teleconferencing may well influence management and thus improve business efficiency.

2.3.3.2 Types of Teleconferencing Service Available (ref 14)

(a) Conference Calling by Telephone: Conferees use telephone sets interconnected via a conference bridge such that any speaking conferee can be heard by all the others. Typically, one exchange line is used plus a number of extensions of each terminal are connected. The conference can be held without anyone moving from their desk, also costs are relatively cheap. However, no speaker recognition is given automatically, the speech is relatively poor and hands-free conversation is difficult.

(b) Simple Loudspeaker System: This method uses loudspeaking telephone units at each terminal which provides hands-free conversation. More than two people may confer at each end provided they share one unit. Voice switching is used to introduce loss on the return speech to minimise acoustic feedback. No acoustic treatment to the rooms is required and the units are simple to use but the speech quality is reduced by clipping and again no automatic speaker identification is given.
(c) Audio-Visual Systems: These include the 'Bell picture phone' and British Telecom Confravision. The first system permits three people assembled within view of a TV camera and are viewed on a small telephone unit at the remote end. The limited bandwidth allows only low resolution pictures of facial detail and expression. A more sophisticated version incorporates voice switching to interconnect several picture phones. Each conferee has a TV display on which the face of the current speaker is viewed on the screen by all other conferees and the current speaker at the same time views the face of the previous speaker. The disadvantages are its high cost and complexity.

For the British Telecom system, up to five conferees are allowed at each of two pre-arranged locations and communicate via a 2-way broadcast standard 625-line vision plus sound channel (TV standby network). Voice switching is not required and the high fidelity sound is full duplex. The conferees appear continuously on TV screens and this provides adequate indication of facial expression with implicit speaker identification. The disadvantages are that the conferences are based on fixed locations and businesses must book in advance and also travel to the nearest confravision studio. The high costs are not likely to attract a large number of business organisations.

Hitherto, there is no system which will satisfy the user requirements of being effective and economical.
There therefore seems to exist a potential for audio teleconferencing upon which marketing departments are endeavouring to forecast a projection of anticipated service. Research work suggests that the value of face-to-face subtleties is unlikely to influence the outcome of a meeting. The special cases of counselling, bargaining and getting-to-know people favour face-to-face meetings. It is estimated that some 53% of all meetings in the UK by 1990 could be carried out by audio teleconferences.

2.3.3.3 Fundamental Requirements for Audio Teleconferencing

As far as possible, it is required to make the teleconferencing equipment as 'transparent' as possible to the users such that it neither intrudes or inhibits the users' normal behaviour. It is suggested that the basic requirements for any audio teleconferencing should be;

a) **Loudness**

Adequate loudness of the received speech from the distant terminal is required. If the signal is too quiet then the system is extremely tiring and frustrating to use. The desired signal level depends upon ambient noise, the room acoustics and the sharpness of hearing. The most unfavourable condition occurs when a quiet room is connected to a noisy one via a telephone link. A method of rebalancing the conferencing system gains to adapt to the ambient noise in the terminal rooms is clearly desirable.
b) Bandwidth and Speech Quality

Ideally, a bandwidth in excess of 10kHz should be provided if a high-quality audio-teleconference is required. Most private currents and the public switched telephone network (PSTN) are limited to the telephone bandwidth. If the speech is restricted to the telephone bandwidth it should be presented at the terminal at 3-6dB greater than if it were broadband. If this extra loudness is available then conferees presented with a properly optimised telephone bandwidth are normally unable to recognise that the system is quite severely bandlimited, unless they are deliberately allowed to listen to a wide band system and make a direct comparison. The quality of speech produced by the receiver in a telephone handset is widely and wrongly taken as the quality to be expected from a telephone network; users may be surprised by the network performance revealed by the electro-acoustic devices forming the audio-teleconference system.

c) Howl Prevention

To prevent howling caused by acoustic feedback instability, voiced-activated switched attenuation is extensively used in loud speaking telephones and audio teleconferencing units. Also to prevent howling, a 5Hz frequency shifter is used. The frequency response of the complete audio loop without the frequency shifter is very irregular caused mainly by interference effects from the many sound modes in the terminal rooms. As the loop gain is increased
the system oscillates at the frequency where the largest peak in the response occurs. The frequency shifter capitalises on the rapid fluctuation of the frequency characteristics and attempts to reduce or cancel a peak by merging it into an adjacent trough. The frequency shifter allows a small amount of extra gain to be added to the audio loop before colouration of the speech ensues.

More current methods of howl prevention are in the direction of echo cancellation devices (ref 10).

Having discussed audio teleconferencing systems in general, two distinct systems are briefly described.

2.3.3.4 'Orator' - The British Telecom Audio Teleconferencing System

Historically, the design of the ORATOR system followed several stages of evolution. Initially, the terminals would be connected via a four-wire private circuit. Once the design of the basic terminal had been established the PSTN operation was then developed.

The acoustics design employs commercially available microphones and loudspeakers as shown in figure 2.16. This arrangement ensures even pick up of speech from conferees seated around the table. In practice, the system gain is set up by using an equalised artificial voice emitting a speech weighted noise spectrum. A room noise adaptation circuit is also included which adjusts the ratio of send to receive gain to compensate for room noise whilst keeping the terminal loop gain constant.
The electronics associated with the early experimental designs of the conference terminal which relied on a four-wire private circuit for interconnection, consisted of microphone pre-amplifiers, a mixer, a limiter, a lowpass filter to restrict the transmitted speech to the telephone bandwidth, a line driver amplifier, a receive amplifier and a power amplifier to drive the loudspeaker. To these were added a shallow voice switch and a noise adaptation circuit. Greater flexibility is possible if the system will operate over the PSTN. The major problem with the PSTN is that the loss of any connection that is established is unknown. The audio teleconference terminals require a circuit of known and repeatable loss in order to preserve loop stability and maintain an adequate received speech level, thus an adaptive system is required to equalise the loss of the PSTN connections. The electronics unit uses a microprocessor to control the set up and equalisation of the connections. Equalisation is made of both the overall loss and frequency response of the lines using two tones 700Hz and 2800Hz. Equalisation of the PSTN circuit is initiated by the user. The problems of line noise are overcome by including a 2 to 1 logarithmic compander. The lower cut-off frequency of the speech band filter minimises the effects of strowger switching or other unwanted line noise.

Both the locations possess similar terminal equipment acoustics unit and electronics unit and are connected by 2 telephone lines (4 wires) to ease the loss equalisation problem. More currently, a multi-location system is being developed and further voice switching is required to reduce the causes of feedback howl.
2.3.3.5 The Plessey Remote Conference System (ref 14)

This system differs very little in design from the British Telecom Orator system; it is also aimed at a similar market. For the Plessey system, each conference site is arranged for up to 8 people (16 person joint conference). Each position has a directional microphone, a loudspeaker unit plus an illuminated nameplate, the nameplates being visible to all conferees to facilitate automatic speaker recognition. The level of sound input must exceed a preset threshold before the notional loudspeaker is activated and the nameplate is illuminated at the distant end. A conferee has then 'captured' the system which is indicated by a capture-lamp associated to him. In the capture mode, all the background sounds are heard from the loudspeaker in the capture position.

The speech and control signalling is provided by a 4 wire private telephone bandwidth circuit linking the two sites. It is only necessary to indicate a change of state when a conferee captures the system. Thus only a low data rate is required and it is continually updated to guard against noise to protect against errors. The data tones are transmitted in an exclusive spectral slot in the speech band. Data modulation is in the form of two-state frequency shift keying (FSK) at 1750kHz and 1850Hz equivalent to a 100 Baud rate. The chosen data tone frequencies fall within the range where the transmission line characteristics are well defined for more reliable operation. Speaker identification is encoded at the transmitter, filtered from the spectral slot and decoded at the receiver.
The Plessey arrangement is considered to provide operating conditions which approximate to an actual face-to-face round the table conference. It does have the slight refinement over the British Telecom system in that speaker identification is provided at the expense of an increase in equipment complexity.

The audio teleconference systems described pose themselves to a direct application of wideband speech compression and expansion. If wideband speech can be effectively provided at each terminal but using a telephone bandwidth link then teleconferences can be made to approximate more closely to a face-to-face conference without any increase in transmission costs (i.e. telephone line charges) to the participating business.

2.3.4 'Phone-In' Contributions to Broadcast Radio and TV Programmes

The advent of commercial radio in the United Kingdom has firmly established a previously little known broadcasting notion, namely the phone-in programme (ref 15). As the interest in audience participation has grown, British Telecom has attempted to cater for the high level of calls involved during a phone-in programme. As far as the public switched telephone network is concerned, the main problem lies in the ratio of calls made to those answered which inevitably causes congestion to the detriment of other telephone users. British Telecom now developed the network such that the telephone switching equipment at various stages of the routing of a call would recognise it and limit the total calls extended to the
radio or TV station. In this way the funnel of calls which previously caused much damage and congestion was replaced by a steady stream of calls which would be answered thus preventing possible frustration and nuisance for many other customers.

As far as voice-communication is concerned, the problem here is the disparity of speech quality exhibited during a phone-in conversation between the interleaved speech of a calling subscriber and that of the studio presenter. To the radio listener, the continual change in mental concentration required to listen to telephone quality speech and studio quality speech becomes very tiring and annoying. At present, neither British Telecom nor the broadcasting stations have provided any equipment to alleviate this speech quality disparity.

What is now briefly described (ref 16) is the procedure adopted when a phone-in programme takes place and the associated interfacing equipment currently used to connect a telephone line to the broadcasting studio.

When a subscriber dials-in to the radio or television station, the call causes a lamp to glow on a 'key-and-lamp' unit situated in the control room at the station. The call is then answered by an operator who will acknowledge with the name of the broadcasting station and after exchanging relevant details and instructing the caller to switch off his radio set, the call is then held for queuing to talk to the presenter. When the presenter is ready to
accept the call, the operator can route the call in the studio room via the 'telephone balancing unit'. The presenter or engineer can manually adjust the level of the subscriber's voice as the conversation ensues. The subscriber will now be talking to the presenter in 'real-time', i.e. live, but the signal comprising the presenter's speech and the subscriber's speech will be broadcast via a delay line of between 1 and 10 seconds (the Broadcasting Authority's statutory requirements). This delay line formerly implemented by a tape recorded arrangement is now implemented as solid state hardware which stores two 20kHz stereophonic channels.

At the start of a phone-in programme, the broadcast signal needs to be changed from real-time to delayed transmission; the delay is gradually built up by lengthening the pause intervals between the speech signal thus preventing any discontinuity in the broadcast signal. The reason for using the delayed broadcast signal is to safeguard against any offensive material from the calling subscriber being broadcasted. At such a time the presenter or producer will activate the 'censor-button' which will revert the programme back into real-time and the memory data of the hardware storage is dumped. Upon release of the censor-button, the programme-delay builds up as before.

When the subscriber's contribution is finished, the subscriber clears down and the operator routes the next call to the presenter.

All the calls are routed via the telephone balancing unit; the object of this unit is to enable a telephone subscriber to listen to the signal from the broadcasting station via a standard British
Telecom subscriber line and at the same time for the sound balance engineer to mix into the station output a signal from the telephone subscriber with minimal interference from the station's own feed to that subscriber. Ideally, the feed from the subscriber should be completely devoid of the 'send' signal but due to practical considerations this situation is not achieved.

The telephone balancing unit aims to alleviate this problem in two ways:

(1) **Hybrid balancing:** The output stage of the balancing unit - an electronic analogue of the 2-4 wire hybrid transformer with the addition of compensation components for cancellation of capacitive elements of the line.

(2) **Comping:** The station return signal from the subscriber's line passes through as expander/compressor activated from the incoming clean station feed. The effect of this is to provide blocking of the station return signal and to improve the signal to noise ratio of the subscriber's signal.

The system diagram of the telephone balancing unit is shown in figure 2.17. This shows the connections of the subscriber's line, the presenter's speech from the mix and subscriber's speech to the mixer. The other features shown are the suppression of the presenter's echoed speech signal from the B.T. line. If the reactive and resistive compensation potential dividers are adjusted
correctly then phase cancellation occurs. The other salient feature of the telephone balancing unit is the gain control of the subscriber's return signal. The gain is inversely related to the level of the presenter's speech. Thus when the presenter actually is talking, this has the effect of 'ducking' or blocking the subscriber's speech sent to the mixer.

There are no speech quality improvement devices incorporated within the balancing unit; it remains a motivation of this project to attempt to provide such a possible means. This topic will again be dealt with in Chapter 5 of this thesis.
3.1 Introduction

Speech sounds consist of rapid fluctuations in air pressure which give rise to acoustic waves. The transmission of the acoustic wave is only feasible over short distances, hence it is not a good means for distant transmission. With electrical means of transmission, a substantial mismatch exists between the capacity of the human source-sink and the capacity of the waveform channel; the channel is capable of transmitting information at rates much higher than the information rate that the human can assimilate. A more efficient means for speech coding and also for reducing the bandwidth used to transmit the speech is constantly being investigated.

The written equivalent of the information rate generated by the human is less than 50 bits per second. The maximum rate of transmission for an arbitrarily small error rate is \( C = B \log_2 \left( 1 + \frac{S}{N} \right) \) bits per second. A conventional waveform voice channel has a bandwidth typically around 3000Hz and a signal to noise ratio of about 30dB. The formula therefore indicates that such a channel has the capacity to transmit information at rates on the order of 30000 bits/sec. Similar bit rates are encountered in conventional PCM where the signal is sampled at the Nyquist rate of twice the
signal bandwidth. For a 64 level (6 bit) quantization, a typical channel capacity would be $2x(3000)x\log_2 64 = 36000$ bits/sec. Thus the channel capacities are of the order of 600 times greater than that apparently required for the written equivalent, i.e. acoustic speech appears to require 600 times more information capacity. This large disparity suggests that the full information capacity of an audio channel is not necessary for speech transmission. The human listener however, can be very selective (ref 17) in deciding what aspects of the signal are chosen for attention i.e. voice quality of the speaker, background noise or even to the way certain speech sounds are produced (i.e. the 'tone' of the voice). The written equivalent of the message might be largely unnoticed. Any sounds that do not fit in with the expected quality of the signal will attract the listener's attention.

The objective of designing speech coders is to be able to transmit speech with the highest possible quality over the least possible channel capacity with the least cost. These three attributes are mutually conflicting, thus the design reduces to finding the appropriate compromise to suit the intended application. Once the coder is designed, it is then categorised into an operating digit rate (or bandwidth); then the recovered speech performance is assessed by some form of fidelity criteria or subjective listening tests. This is to develop a means of comparison of other designs of coders operating at similar digit rates.
The speech coder designs reviewed in this chapter are broadly divided into three main areas; namely analysis/synthesis systems (vocoders), waveform coders and intermediates. The latter is a class which combines properties of the two former classes. Their complexity is greater than for simple waveform coders and some of the higher performance systems may be more complicated than vocoders.

3.2 Waveform Coders

Waveform coders, as their name implies, attempt to duplicate the actual shape of time waveform produced by the microphone and its associated analogue circuits. In principle, they are designed to be signal-independent, hence they can code equally well a variety of signals, e.g. speech, music, tones, voiceband data and video signals. They also tend to be robust for a wide range of talker characteristics and for noisy environments. To preserve these advantages with minimal complexity, waveform coders typically aim for moderate economies in transmission bit rate.

In nearly all waveform coding systems, the analogue signal is quantized in both time and amplitude. Quantization in time means that the analogue signal is sampled at certain instants and the transmitted data is related only to these samples. If the bandwidth of the signal is limited, the sampling theorem states that it is theoretically possible to reconstruct the waveform if the sampling frequency is taken at least twice the bandwidth. Quantization in amplitude means that the continuous amplitude range of input samples
is replaced by a set of finite number of discrete amplitude levels. This inherently introduces an error in the amplitude of the samples, known as quantization noise.

3.2.1 Pulse Code Modulation

The significance of pulse code modulation is that, historically, it is the first method (suggested by Reeves (ref 18) in 1938) for converting analogue speech signals into digital form and is still widely used for feeding analogue signals into computers or other digital equipment for subsequent processing (in which case it is known as analogue-to-digital conversion [ADC]).

The processes involved in a PCM codec are described in great detail by Cattermole and are as follows:

The input speech signal $x(t)$ is band limited to exclude any frequencies greater than $f_{\text{max}}$. This signal is sampled at a rate $F_s$ equal to or greater than the Nyquist rate $2f_{\text{max}}$. The samples are then quantized into one of $2^B$ levels. This implies an information of $B$ bits per sample, and an overall information rate of $F_sB$ bits per second (b/s). The discrete amplitude levels are represented by distinct binary words of length $B$. For example, with $B=2$, one can represent 4 distinct levels using the code words 00, 01, 10 and 11. For decoding, the binary words are mapped back into amplitude levels, and the amplitude time pulse sequence is lowpass filtered with a filter whose cut off frequency is $F_s/2$ to reproduce the analogue decoded signal $\hat{x}(t)$. 
3.2.1.1 Time Invariant Quantizers (refs 19, 20)

The quantizer is the element which in PCM determines the accuracy of the approximation of the recovered signal \( \hat{x}(t) \) to the input signal \( x(t) \), assuming no transmission bit errors. In its simplest form it is called the zero memory or memoryless quantizer. A zero memory quantizer accepts analogue signals and imposes amplitude restrictions on them so that each analogue sample is forced, i.e. quantized to the nearest of a finite set of amplitude levels. Consequently, the value of the quantized sample is independent of earlier analogue samples applied to it.

An \( n \)-level zero-memory quantizer is defined by the set of \( n-1 \) decision levels \( x_1, x_2, \ldots, x_{n-1} \) and a set of \( n \) output levels \( y_1, y_2, \ldots, y_n \). When the input sample 'x' lies in the \( i \)'th quantization interval it is quantized to a value which is contained within the interval.

The input-output characteristic of a zero-memory quantizer can assume differing symmetries about the zero level as shown in figures 3.1(a) and 3.1(b). They can be viewed as a stair-case approximation of the input sample 'x'.

Let the quantum step size be denoted by \( \Delta \) where \( \Delta = x_i - x_{i-1} \) i.e. the spacing between uniform quantization levels. The quantization noise (or error) introduced is bounded and is sometimes
known as granular noise. If the number of quantum levels is large, the quantization error, $E$, has the following uniform distribution:

$$ p(E) = \frac{1}{\Delta} , \quad -\frac{\Delta}{2} \leq E \leq \frac{\Delta}{2} $$  \hspace{1cm} (3.1)

This is not true if the signal saturates the quantizer. The mean-square value of the quantization error is:

$$ N_q^2 = \int_{-\Delta/2}^{\Delta/2} E^2 p(E) \, dE = \frac{\Delta^2}{12} $$  \hspace{1cm} (3.2)

If the rms value of input $x$ is $\sigma_x$ then the signal-to-error ratio (SNR) is

$$ \text{SNR} = \frac{\sigma_x^2}{N_q^2} = \frac{\sigma_x^2}{\left[\frac{\Delta^2}{12}\right]} $$  \hspace{1cm} (3.3)

Note that for a given step size, the SNR is dependent upon the variance of the input signal.

Let the quantizer include the amplitude range $-4\sigma_x$ and $+4\sigma_x$, if the signal PDF is modelled by a zero mean Gaussian function, signal samples will fall outside the $8\sigma_x$ quantizer with a probability (of overload) of $< 10^{-4}$.

The quantum step size is now equal to $8\sigma_x/2^B$. From equations (3.2) and (3.3)

$$ \text{SNR(dB)} = 10 \log_{10} \text{SNR} = 6B - 7.2 $$  \hspace{1cm} (3.4)
Equation (3.4) shows that the SNR of a $2^B$ level quantizer increases linearly with the number of bits $B$. However, the bandwidth of the transmitted bit stream also increases proportionally with $B$. The SNR formulation implies that quantization error samples can be modelled as additive noise samples provided the quantization is sufficiently fine. For coarse quantization (say $B < 5$) the error waveform has too much structure and too much correlation with the input speech itself to be regarded as additive noise.

Non-uniform quantization is characterised by fine quantizing steps (and hence, a relatively small noise variance) for the very frequently occurring low amplitudes in speech; while much coarser quantizing steps take care of the occasional large amplitude excursions in the speech waveform. Average distributions of speech amplitudes are decreasing functions of amplitude and non-uniform quantization constitutes a direct utilisation of this speech property. A quantizer characteristic that is equivalent to the uniform quantization of a logarithmically compressed input is used in commercial telephony. The amplitude compression characteristics used in log-quantization follow either the so called $\mu$-law or the A-law (invented by Cattermole)\(^{(\text{ref 21})}\). For a signal input $x$, both characteristics are symmetrical about $x=0$. For $x > 0$ and $x_{\text{max}}=1$, compressed signals $x_\mu$ are defined as follows:

\[
\mu\text{-law: } x_\mu = \left[ \ln(1 + \mu x)/\ln(1 + \mu) \right], \quad \mu > 0 \quad (3.5)
\]

\[
\text{A-law: } x_\mu = \frac{Ax}{(1 + \ln A)}, \quad 0 \leq x \leq A^{-1} \quad (3.6)
\]

\[= \frac{(1 + \ln Ax)/(1 + \ln A)}{A^{-1} \leq x \leq 1}\]
A commonly used value for the parameter $\mu$ is 255 and 'A' the compression parameter takes values close to 86 for a 7-bit speech quantizer.

It was mentioned in equation (3.4) for a uniform quantizer that the SNR was dependent upon the signal variance; in order to render the SNR to be signal invariant, it can be shown (ref 21) that a truly logarithmic compression needs to be employed. Unfortunately, a pure logarithmic function does not pass through the origin but this is obviated by using a linear approximation for low levels of signal and logarithmic characteristic for high levels. For both the 'A' and 'µ' law characteristics, the SNR can be close to that of a uniform quantizer and remain relatively constant over a wide range of input power. This means that, for a specified dynamic range, these companded quantizers offer a reduction in the number of bits per sample required by a uniform quantizer to accommodate the same dynamic range of input signals.

In typical voice communication systems, the dynamic range of speech signals can be as much as 40dB. While time-invariant non-uniform quantization has been a traditional solution to this problem, better results can be obtained by recognising that the large dynamic range of speech signals is a result of a non stationary or a time-varying process at the coder input; so that a truly optimal quantization strategy is one that is also time-variable or adaptive to the input signal.
3.2.1.2 Adaptive Quantization

Adaptive quantization utilises a quantizer characteristic (uniform or non-uniform) that shrinks or expands in time like an accordion. Snapshots of such an adaptive quantizer at two instants of time may therefore look like the pictures in figure 3.2(a) and (b) indicating adaptation for low and high speech power levels respectively. Although speech signals have a large dynamic range over a long period of time, input power levels vary slowly enough to facilitate the design of simple adaptation algorithms to keep track of these power variations.

An efficient way of matching the quantizer's step size to the signal's variance is the "One-Word-Memory" adaptive quantization suggested by Flanagan and developed at length by Jayant. (ref 22) Consider at the \( i \)th sampling instant, the step size of a B bit uniform quantizer to be \( \Delta_i \) and its output level \( x_i \), i.e.

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

(3.7)

At each sampling instant the step size \( \Delta \) is multiplied by a fixed expansion-compression coefficient which is determined from the previous quantizer output level. Thus at \((i+1)\)th instant the value of the step size \( \Delta \) is

\[
\Delta_{i+1} = \Delta_i \cdot M_\xi[|I_i|]
\]

(3.8)

where \( M_\xi[\cdot] \) is a function of 'i'.
When the multiplier function is properly designed, the adaptation logic serves to match the step size to an updated estimate of the signal variance. $M_2$ is one of the $l$ fixed coefficients corresponding to the quantizer's output levels. When $B$ is even the number of coefficients is $B/2$ while for $B$ odd there are $(B+1)/2$ coefficients. The $M_2$ coefficients are less than, but close to unity for coefficients corresponding to the inner quantization levels. Values between 1 and 2.5 are used for the outer levels of the quantizer. With this strategy the rate at which the step size $\Delta$ is increasing is greater than its rate of decrease and the occurrence of possible overload errors (which have a serious impairment as far as perception is concerned) is minimised.

In the static operation the amplitude range of the quantizer matches the $\sigma_x$ value of the incoming input sequence and the $M_2$ coefficients must be such that the step size $\Delta$ tends to its optimum value to produce the maximum SNR. The dynamic behaviour of the quantizer is related in the speed the step size $\Delta$ can adapt to sudden large changes of the input level, and depends upon how close to, or far from unity are the $M_2$ levels of the inner and outer quantization levels respectively.

3.2.1.3 Dithered Quantization

In this case, a pseudo-random noise sequence is added to the speech that is to be quantized; subsequent subtraction of the pseudo-random sequence from the quantizer output provides a white-noise-like error
sequence (which is found to be less objectionable, perceptually, than the signal dependent error waveform) and also preserves the original value of SNR. The technique is most useful in the narrow but practically significant range $4 \leq B \leq 6$. Specifically, dithered quantization noise seems to mask consonant sounds more than straightforward quantization error. Dither can only be applied successfully to fixed level quantizers, as adaptive quantization technique and especially instantaneous ones, tend to produce a signal independent error pattern.

3.2.2 Differential Coding Systems

Speech that is sampled at the Nyquist rate exhibits a significant correlation $C_i$ between successive samples. One consequence is that the variance of the first difference

$$D_i(1) = x_i - x_{i-1}$$

is smaller than the variance of the speech signal itself

$$\langle D_i^2(1) \rangle = \langle (x_i - x_{i-1})^2 \rangle$$

$$= \langle x_i^2 \rangle + \langle x_{i-1}^2 \rangle - 2 \langle x_i x_{i-1} \rangle$$

$$= \langle x_i^2 \rangle \cdot 2(1 - C_i)$$

(3.10)

where $C_i$ is the correlation between adjacent samples. Obviously, if $C_i$ is greater than 0.5, $D(1)$ has a smaller variance than $\{x_i\}$. As a result, it is advantageous to quantize $D(1)$ instead of $\{x_i\}$ and
use an integrator to reconstruct \( \{x_i\} \) from the quantized values of \( D(1) \). This is because for a given fineness of quantization \( B \), the quantization error power is proportional to the variance of the quantizer input. More generally, if the quantizer input is \( x_i - a_1 x_{i-1} \), it can be shown that the variance of this quantity is minimum for \( a_1 = c_1 \). Formally the quantity \( a_1 x_{i-1} \) is called a first-order prediction of \( x_i \) and the differential coding scheme is called predictive coding. A more general (linear) predictor would be one of order \( p \), where the estimate of the input signal sample \( x_i \) is

\[
\hat{x}_i = \sum_{r=1}^{p} a_r x_{i-r}
\]  

(3.11)

The values of the prediction coefficients \( a_r \) are calculated in terms of the first \( p \) samples of the autocorrelation function of \( \{x_i\} \). The error sequence (which is quantized) is taken as \( \{e_i\} = \{x_i\} - \{\hat{x}_i\} \).

To illustrate the advantage of differential encoding over straightforward quantization, consider \( M \) input samples to be PCM encoded and transmitted with a total of \( M \cdot N \) bits. Consider also the same \( M \) samples to be differentially encoded so \( M \) error samples are quantized with \( N_1 \) bits/sample accuracy and transmitted together with \( N_2 \) bits of information related to the \( \hat{x}_i \) estimation procedure parameters. The variance of the error sequence is much smaller than that of the original speech samples and the bits per sample needed to describe, with the same accuracy as in PCM, the \( e_i \) samples are less than \( N \) i.e. \( N_1 < N \). Generally, \( N_2 < < N \) and
therefore $M \cdot N > (M \cdot N_1 + N_2)$. Thus the main characteristics and objective of differential encoding is the considerably smaller amplitude range of the error sequence, when compared with the input signal itself.

3.2.2.1 Differential Pulse Code Modulation (DPCM)

Differential Pulse Code Modulation is based upon an invention by C. C. Cutler in 1952, which uses quantization of the differences between the successive Nyquist samples of equation (3.9). Alternatively, $\hat{x}_i$ from equation (3.11) may be used in place of $x_{i-1}$.

At present, although it is well known that DPCM is a more efficient way of encoding speech signals than PCM, the latter is employed almost exclusively in all the commercial digital transmission systems. The reasons for this choice are:

(i) At the beginning of the sixties, PCM was established as a viable method of digital communications; consequently large capital investment was made in the direction of PCM terminal equipment. DPCM at that time was still being investigated.

(ii) Also at that time, compromise predictors and fast adaptive predictors had not been developed; the dependence of the DPCM performance upon the statistics of the input signal appeared to be a serious weakness particularly in the case of the telecommunications networks which have to convey signals
other than speech. When the long-term statistics of the input signal are different than those used in the design of DPCM, the system may lose its encoding advantage over PCM unless a compromise or an adaptive predictor is employed.

More recently, however, the CCITT has been involved in producing a recommendation for ADPCM (Adaptive Differential Pulse Code Modulation) using a 32 kbits/s transmission bit rate. It is envisaged that this will replace the 64 kbits/s log-PCM that is currently used.

The schematic diagram of the DPCM codec is illustrated in figure 3.3. 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 the linear predictor in the feedback loop of the encoder, based upon previously decoded samples, provides a sequence $\{\hat{x}_i\}$ of predicted samples and an error sequence $\{e_i\}$ is produced whose $i^{th}$ element is

$$e_i = x_i - \hat{x}_i \quad (3.12)$$

The error samples are quantized to produce $\{q_i\}$ by the quantization process. The samples at the output of the quantizer are then binary encoded and transmitted as well as being locally decoded in the feedback loop of the encoder.
The quantizer is included inside the predictive closed loop such that the quantization noise associated with the reconstructed sequence \( \{ \hat{x}_i \} \) is the same as that of the error sequence \( \{ e_i \} \), i.e. \( \{ q_i \} \). This can be seen from the following equations, applicable to the \( i^{th} \) sampling instant.

\[
\hat{x}_i = e'_i + \hat{x}_i
\]

\[
e'_i = e_i + q_i
\]

\[
e_i = x_i - \hat{x}_i
\]

\[
\therefore \hat{x}_i = x_i + q_i
\]

On the other hand, if the quantizer is placed outside the feedback loop, there is an accumulation of quantization noise at the output of the decoder.

3.2.2.2 Delta Modulation

Delta modulation (DM) may be considered perhaps as the simplest form of DPCM. Quantization of the error signal is to a one-bit accuracy only (i.e. a simple comparator). A single or double integrator is typically used as the predictor network as shown in figure 3.4. The transmitted binary samples, \( e'_i \), are either +1 or -1 and represent the sign of the error \( e(t) \). The integrator can be implemented in various ways, including a simple analogue storage capacitor. The band limited input signal (bandwidth \( W \)) is sampled at a rate \( F_s \).
which is much higher than the Nyquist frequency, and a staircase approximate to the input is constructed. The local estimate provided by the integrator \( \hat{y}(t) \) is shown in figure 3.5. The step size of the staircase function is determined by the amplifier constant, \( k \), which is typically chosen to be small compared to the input signal magnitude. Two types of distortion can occur in the estimate, granular distortion and slope overload. The former is determined by the step size of the quantization (i.e. the amplifier \( k \)). 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.

\[
|\dot{x}| > k/T \tag{3.13}
\]

These two types of distortion are indicated in figure 3.5.

Granular distortion can be made small by using a small step size, slope overload distortion can be reduced by using a large step size or by running the sampler clock faster. The latter however increases the transmission bit rate. For a given bit rate, the step size is selected to give a compromise between quantizing distortion and slope overload. Perceptually, more overload noise power is tolerable than granular noise power (ref 23); during granular distortion the samples of the error signal tend to be uncorrelated and the error signal power spectrum tends to be uniform.

For high-quality speech transmission, say with SNR's of the order of 40dB, the resulting bit rate for simple DM is relatively high,
typically greater than 200 kbits/s. There is thus strong interest in techniques for reducing the bit rate of DM while at the same time retaining most of the advantages of circuit simplicity. Adaptive delta modulation is one such solution.

A) Adaptive Delta Modulation (ADM)

In ADM, the quantizer step size is varied according to a prescribed logic which is chosen to minimise quantizing and slope distortion when the sampler is run at a relatively slow rate. This additional control is normally provided by a step size multiplier incorporated in the feed-back loop, as shown in figure 3.6. The step size control logic may be discrete or continuous and it may act with a very short time constant (i.e. sample-by-sample) or with a time constant of syllabic duration. Normally, the step size is controlled by information contained in the transmitted bit stream, but it may be controlled by some feature of the input signal, e.g. the slope magnitude averaged over several msec. (ref 24) In this case, the control feature must be transmitted explicitly along with the binary error signal as side information.

The receiver duplicates the feedback (predictor) branch of the transmitter, including an identical step size element and in the absence of errors in the transmission channel, the receiver duplicates the transmitter's estimate of the input signal. Desampling by a lowpass filter to the original bandwidth completes the detection.
B) Constant Factor Delta Modulation (CFDM)

One form of adaptive delta modulation is constant factor delta modulation (Jayant). (ref 25) The encoder responds to the instantaneous variations in the analogue signal and is suitable for encoding both speech and television signals. The encoder is shown in figure 3.7. The output binary waveform $L(t)$ is transmitted to the receiver as well as being connected to the adaptation logic in the feedback network. This logic has a one bit digital delay $D$ so that it can inspect the binary level of the $L(t)$ signal at say sampling instant $i$ and also the previous sampling instant $(i-1)$, the corresponding values of $L(t)$ are $L_i$ and $L_{i-1}$, respectively.

If $L_i$ and $L_{i-1}$ are both ones or zeros (or minus ones) it indicates that the error has not changed sign for two successive clock periods and some increase in the step size of the feedback signal to the subtractor is warranted. On the other hand if $L_i$ and $L_{i-1}$ are different the error has changed polarity between successive clock periods. This suggests that the step change in $y(t)$ should be reduced in magnitude. In figure 3.7, if the output from the Exclusive-OR is a logic one the switch is moved to the negative voltage source ($-B$). If the output from the Exclusive-OR is a logic zero the switch connects the positive voltage ($+A$) to the input of the multiplier.
At a particular sampling instant, \( i \)

\[
Z_i = \begin{cases} 
\text{impulse } A \text{ if } L_i = L_{i-1} \\
\text{impulse } -B \text{ if } L_i \neq L_{i-1}
\end{cases}
\]  

(3.14)

The final waveform \( y(t) \) feedback to the error point is the integral of \( m(t) \), where \( m(t) \) is the waveform at the output of the multiplier. At the \( i \)th sampling instant \( m(t) \) is \( m_i \) and is formed by multiplying the previous value of \( m_i \) by \( z_i \).

\[
m_i = z_i m_{i-1}
\]  

(3.15)

The production of \( m_{i-1} \) requires the use of an analogue delay of one bit duration.

\( z_i \) is an impulse and consequently \( m_i \) is also an impulse with the result that the waveform \( y(t) \) contains steps of varying height. In particular, if \( y_i \) is the value of \( y(t) \) at the \( i \)th sampling instant, then

\[
y_i = y_{i-1} + m_i
\]  

(3.16)

This is because \( m_i \) is integrated to give the new value of \( y(t) \), namely \( y_i \). The integral of an impulse of strength \( m_i \) is a step of height \( m_i \). Substituting for \( m_i \)

\[
\Delta y_i = y_i - y_{i-1} = z_i m_{i-1}
\]  

(3.17)
As $z_i$ is either A or -B, the ratio between two successive steps in the feedback waveform differ by a constant factor $z_i$. It is because of this important property that this type of encoder is called a constant factor delta modulator.

The local decoder therefore operates on $L_i$ and $L_{i-1}$ to produce changes in $y_i$ which attempt to follow $x_i$ and minimise $e_i$. The forward path of the encoder undertakes the usual task of producing an output impulse $L_i$ whose polarity is identical to the polarity $e_i$ at the $i^{th}$ sampling instant.

The decoder has the usual arrangement of a local decoder followed by a lowpass filter to remove unwanted noise.

C) Continuously Variable Slope Delta Modulation (CVSD) (ref 26)

Continuously Variable Slope Delta Modulation, also known as Syllabically Companded Delta Modulation, is a method of compressing large amplitude levels in the signals relative to the smaller ones. This design is used to overcome the problem of having just one input level to maximise the signal-to-noise ratio. The companding is accomplished solely by the delta modulator and therefore has a non-linear active network in its feedback loop. The type of expander which must be employed to recover the original signal at the decoder is to be found in the feedback loop of the encoder. The companded encoder is adaptive in that it adapts the binary signal at the output of the encoder to produce a small error signal. In so
doing the encoder 'expands' the amplitudes of the base band signals which reside in the latent form in the output binary signal. The adaptive algorithms used by Jayant's delta modulators adjust the magnitude of the pulses fed back to the error point at a much slower rate than the instantaneous variations in the speech signal. This rate is, generally speaking, the pitch rate of the speech signal, i.e. of the order of 10mS, and not the syllabic rate which is generally in excess of 100mS.

The CVSD algorithm exploits the syllabic characteristics of the speech waveform to minimise the number of bits required for transmission. Figure 3.9 shows the CVSD algorithm; the general relations are given by

\[ \hat{x}_{i-1} = \alpha \hat{x}_i + |(1-\alpha)\Delta_i| e_i \]  \hspace{1cm} (3.18)

where 

\[ e_i = \text{signum} \left[ x_i - \hat{x}_i \right] \]  \hspace{1cm} (3.19)

and

\[ \hat{x}_i \] is the estimate of the incoming analogue signal

\[ \alpha \] is the leakage factor associated with the estimate integrator.

\[ \Delta_i \] is the \( i \)th step size and

\[ x_i \] is the \( i \)th input sample.

Furthermore, \( \Delta_i \) is generated by syllabic companding and is given by

\[ \Delta_{i+1} = \beta \Delta_i + (1-\beta)(V+V_i) \]  \hspace{1cm} (3.20)
Where \( V \) is constant when three consecutive outputs from the CVSD encoder are identical (sometimes this number can be two or four), \( V_1 \) is another constant voltage added to \( V \) to ensure that the minimum step size is non-zero.

In figure 3.9, the output of the overload detector, \( V \) is either 0 or \( V_2 \) volts depending upon three consecutive digital outputs of the CVSD. The feedback circuit of the encoder is in fact the CVSD decoder.

The CVSD described in reference 27 has a time constant for the step size integrator, \( \tau_1 = 5.69\text{mS} \) and the time constant of the estimate integrator \( \tau_2 = 1\text{mS} \), which gives

\[
\beta = \exp\left( \frac{-1/F_s}{5.69 \times 10^{-3}} \right) \tag{3.21}
\]

\[
\alpha = \exp\left( \frac{-1/F_s}{10^{-3}} \right)
\]

when \( F_s = 16\text{kb/s} \), \( \beta = 0.99 \) and \( \alpha = 0.94 \)

The coefficients \( \alpha \) and \( \beta \) are adjusted differently in different CVSD processors. The constants \( \tau_1 \) and \( \tau_2 \) are chosen with reference to the speech waveform; a typical speech signal has most of its energy residing between 700-1000Hz and has an envelope of 60 to 100Hz. The step size integrator of the CVSD generates the envelope of the speech signal, therefore \( \tau_1 \) is chosen as 5.69mS, which corresponds to approximately 100Hz, and \( \tau_2 \) is adjusted to 1mS, which corresponds to 1000Hz.
3.3 Analysis-Synthesis Systems

In this section, we are concerned with the parametric representation of speech signals; efficient communication suggests transmission of the minimum information necessary to specify a speech event and to evoke a desired response. Implicit is the notion that the message ensemble contains only the sounds of human speech and no other signals are relevant. The basic problem is to design a system so that it transmits with maximum efficiency only the perceptually significant information of speech.

One approach to the goal is to determine the physical characteristics of speech production, perception and language and to incorporate these characteristics into the transmission system. Ideally, the characteristics are described by a few independent parameters, and these parameters serve as the information bearing signals. Transmission systems in which a conscious effort is made to exploit these factors are generally referred to as analysis-synthesis systems.

In an ideal analysis-synthesis system, the analysis and synthesis procedures are presumably accurate models of human speech production. Additional economies in transmission can accrue from perceptual and linguistic factors. The pure analysis-synthesis system therefore has the greatest potential for bandsaving and its analysis and synthesis processing typically require complex operations.
Speech signals can be described in terms of the properties of the signal producing mechanism, that is the vocal tract and its excitation. This characterisation forms the common basis for a large class of bandwidth compression systems, the idea is shown in figure 3.10. The three operations involved are first, the automatic analysis of the signal into quantities that describe the vocal excitation mode and structure, second, the multiplexing and transmission of these parameters, and finally, the reconstruction (synthesis) of the original signal from the parameters.

3.3.1 Channel Vcoders

Analysis-synthesis telephony was initiated with Homer Dudley's invention of the Vocoder (contraction of the words Voice Coder). This has become largely a generic term, commonly applied to analysis-synthesis systems in which the excitation and systems functions are treated separately.

Following the coding scheme illustrated in figure 3.10, the Vocoder incorporates one important constraint of speech production and one of perception. It recognises that the vocal excitation can be broad-spectrum, quasi-harmonic sound (voiced), or a broad-spectrum, random signal (unvoiced). It also recognises that perception, to a large degree, is dependent upon the preservation of the shape of the short-time amplitude spectrum. A block diagram of the spectrum channel vocoder is shown in figure 3.11. (ref 28)
The excitation information is measured by the top branch of the circuit. A frequency discriminator and meter measure the fundamental component of the quasi-periodic voiced sounds. Values of the fundamental frequency and its temporal variations are represented by a proportional voltage from the meter. This 'pitch' signal is smoothed by a 25Hz lowpass filter. Unvoiced sounds normally have insufficient power in the fundamental frequency range to operate the frequency meter. Non-zero outputs of the pitch meter therefore indicate voicing as well as the value of the pitch.

Ten spectrum channels in the lower part of the circuit measure the short-time amplitude spectrum at ten discrete frequency bands. Each channel includes a bandpass filter (300Hz wide), a rectifier and a lowpass filter of 25Hz. The predistorting equaliser pre-emphasises the signal to produce nearly equal average powers in the spectrum-analysing filters. The spectrum-defining channel signals consequently have about the same amplitude ranges and signal-to-noise ratios for transmission. The eleven 25Hz wide signals occupy a total bandwidth of less than 300Hz and must be multiplexed in frequency or time for transmission.

At the receiver, the speech spectrum is reconstructed from the transmitted data. Excitation, 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 outputs of these filters are amplitude modulated by the spectrum defining signals and a short-time spectrum approximating that
measured at the transmitter is recreated. With proper design the synthesised speech can be made surprisingly intelligible. An example of speech transmitted by a 15-channel vocoder is shown by the spectrograms in figure 3.12. The important features such as formant structure and voiced-unvoiced excitation are relatively well preserved.

Since the original development of the vocoder, many different versions and variations have been constructed. The number and spacing of the analysing filters along the frequency scale, their bandwidths, degree of overlap, selectivity and many different pitch extraction and voiced-unvoiced detection circuits have been examined.

Although the intelligibility may be high, practical realisations of conventional channel vocoders generally exhibit a perceptible degradation of speech naturalness and quality. The synthetic speech possesses a machine-like quality which is characteristic of the device. There are several factors which seem to be responsible for this; one is the coding of the excitation data, also voiced-unvoiced discriminations often are made with noticeable errors. Relevant structure in the pitch signal may not be preserved and, under some conditions, octave errors may be made in the automatic pitch extraction. Voiced sounds are synthesised from a pulse source whose waveform and phase spectrum do not reflect certain details and changes of the real glottal waveform. The spectral analysis also has granularity, or lack of resolution, imposed by the number, bandwidth and spacing of the analysing filters. A given speech
formant, for example, might be synthesised with too great a bandwidth. Further, the large dynamic range of the amplitude spectrum may not be covered adequately by practical rectifiers and amplifiers. Also all consonants are excited by a noise source.

The basic channel vocoder design can be improved in several ways; the important excitation problem can be obviated to a large extent by voiced-excitation techniques (discussed later). Also sophisticated pitch extraction methods, such as the cepstrum method (discussed in section 3.3.5) provide more precise pitch and voiced-unvoiced data.

The method used to combine the spectrum channel signals for transmission is either by frequency division multiplexing using quadrature modulation, time division multiplexing or digital transmission.

Although voice quality and naturalness normally suffer in transmission by the vocoder, the intelligibility of the synthesised speech can be maintained relatively high, often with a vocoder having as few as ten channels. For a high-quality microphone input and a fundamental-component pitch extractor, typical syllable intelligibility scores of a ten-channel (250 to 2950Hz) vocoder are on the order of 83 to 85 per cent. Typical intelligibility scores for initial consonants range over the values shown in Table 3.1.
Weak fricatives such as 'th' are not produced well in this system. The 30 per cent error indicated for no initial consonant (i.e. for syllables beginning with vowels) indicates imprecision in the voiced-unvoiced switching. Such syllables were heard as beginning with consonants when in fact they did not start with a consonant.

3.3.1.1 Reducing the Redundancy with Channel Vocoder (ref 3)

Generally, vocoder channel signals are not completely independent and possibilities exist for further processing the signals to orthogonalise them for further reducing the redundancy.

A) Peak-picker

In a vowel sound the entire vocal transmission spectrum is specified by the formant frequencies. Usually, therefore, the neighbouring channel signals in a vocoder are strongly correlated. The peak-picking vocoder attempts to eliminate this dependence; it operates by transmitting three to five channel signals which at any instant represent local maxima of the short-time spectrum. A transmission rate of about 1kb/s is estimated to be required. The peak-picking vocoder is a form of Formant Vocoder, the Formant Vocoder is further discussed in section 3.3.3.

B) Linear Transformation of the Channel Signal

For the n channel signals, a set of m signals, where \( m \leq n \), are formed which are a linear combination of the original n. The
coefficients of the linear transformation constitute an \((m.n)\) matrix of constants. Decoding the \(m\) signals to retrieve an approximation to the original \(n\) is also accomplished by a linear transformation, namely, the transpose of the \((m.n)\) matrix. The coefficients of the transformation are obtained to minimise the mean square difference between the original \(n\) signals and the reconstructed \(n\) signals. For a reduction of \(n=16\) to \(m=6\), it was reported that the output was almost completely understandable, although the quality was substantially less than that of the 16-channel vocoder.

C) Pattern Matching Vocoder

This involves classification of the frequency versus amplitude spectral information of the channel signals into a limited number of discrete patterns. The spectral patterns are associated with phonetic units of speech. The sound analysis is carried out according to a pattern recognition scheme; at any instant, the best match between the short-term speech spectrum and a set of stored spectral patterns is determined. A code representing the matching pattern is signalled to the vocoder synthesizer, along with conventional pitch and voiced-unvoiced data. New information is only signalled when the phonetic pattern changes. At the receiver, a set of spectral amplitude signals approximating to the signalled pattern are applied to the modulators of the synthesizer. The pitch signal supplies the appropriate excitation. Filter circuits are included to provide smooth transitions from one sound pattern to the next.
An early version of the device used a ten channel vocoder and only ten stored patterns shown in figure 3.13. The stored patterns correspond to steady-state spectra of four consonant continuants and six vowels (s, f, r, n and i, I, e, a, o, u) respectively. The intelligibility for common monosyllables is around 50% although the bandwidth required for transmission is only on the order of 50Hz. Its usefulness probably lies in transmitting messages of restricted vocabulary.

3.3.2 Correlation Vocoder

The channel vocoder demonstrates that speech intelligibility, to a large extent, is carried in the shape of the short-time amplitude spectrum. Any equivalent specification of the spectral shape would be expected to convey the same information. One equivalent description of the squared amplitude spectrum is the autocorrelation function. A short-time autocorrelation specification of the speech signal might therefore be expected to be a time-domain equivalent of the channel vocoder.

The correlation vocoder is shown in figure 3.14. In the top branch of the circuit the input speech is submitted to pitch extraction as with the channel vocoder. In the lower branch of the circuit, the input signal is put through a spectral equaliser which in effect takes the square root of the input signal spectrum. This is because the ultimate processed signal is going to be a correlation function whose Fourier transform is the power spectrum (or, the square of the
amplitude spectrum) of the input signal. After spectral square-rooting, the short-time autocorrelation function of the signal is computed for specified delays. This is done by multiplying the appropriate output of a delay line with the input and lowpass filtering the product (the lowpass filter is set to 20Hz). Since the autocorrelation is bandlimited to the same frequency range as the signal itself, the correlation function is completely specified by sampling at the Nyquist interval [i.e. 1/2 BW]. For a 3 kHz signal, therefore, a delay interval $\Delta \tau = 0.167$ms is sufficient. The greatest delay to which the function needs to be specified, practically, is on the order of 3ms. Thus a total of 18 delay channels - each using about 20Hz bandwidth are required. The total bandwidth is therefore 360Hz i.e. about the same as that required by the channel vocoder.

At the synthesizer, voiced sounds are produced by generating a periodic waveform in which the individual pitch period in the correlation function described by the values existing on the $n$ channels. The waveform is generated by letting the pitch pulses of the excitation "sample" the individual $\tau$-channels. The sampling is accomplished by multiplying the excitation with each channel signal. The samples are then assembled in the correct order by a delay line, and are lowpass filtered to yield the continuous correlation function. Since the correlation function is even, the synthesized wave is made symmetrical about the $\tau_0$ sample. This can be done practically with the delay line correctly terminated at its output end, but unterminated and completely reflecting at the far end. Lowpass filtering the samples from the line recovers the continuous signal.
Because a finite delay is used in the analysis, the measured correlation function is truncated and discontinuities will generally exist in the synthesised waveform which leads to a noticeable subjective distortion. The distortion can be reduced by weighting the high-delay correlation values so that they have less influence in the synthesised waveform. The discontinuities are therefore smoothed, and the processed speech obtained approaches that from channel vocoders of the same bandwidth compression.

3.3.3 Formant Vocoders

The formant vocoder aims to encode efficiently the speech in terms of the vocal mode pattern. Formant vocoders generally divide into two groups – namely the cascade and parallel connections of the synthesis circuits. The cascade approach strives to reconstruct the signal by simulating the perceptually significant pole and zero factors of the vocal transmission. The complex frequencies of the poles and zeros and the excitation data (pitch and voiced-unvoiced) are the coding parameters.

The parallel connection attempts to reconstruct the same signal in a different, but equivalent way, namely, from the information on the frequencies of the formants (poles) and their spectral amplitudes (residues). Ideally, the mode frequencies and their residues are specified in complex form. Because the formant vocoder attempts to duplicate the vocal mode structure, it has potential innately for a better and more natural description of the speech spectrum.
The earliest parallel formant vocoder system is illustrated in figure 3.15. At the analyser the input speech band is split into four sub bands. In each band the average rectified-smoothed amplitude, $A$, is measured. These eight parameters which approximate the amplitudes and frequencies of the formants and of voicing are transmitted to the synthesier.

The synthesiser contains excitation circuitry in which three variable resonators are connected in parallel. The fourth branch has a fixed lowpass filter. Voiced (pulse) excitation of the parallel branches is signalled by the voicing amplitude $A_0$. The $A_0$ control also determines the amplitude of the signal passing the fixed lowpass filter. As in the channel vocoder, the frequency of the pulse source is prescribed by $F_0$. Unvoiced (noise) excitation of the parallel branches is determined by amplitude $A_3$. The amplitudes and frequency of the three formant branches are continuously controlled by $F_1$, $F_2$, $F_3$, $A_1$, $A_2$ and $A_3$ and their outputs are combined.

Intelligibility scores reported for the system were approximately 100% for vowel articulation and about 70% for consonant articulation. This is because it is an easier task to analyse the formants of the vowel sounds than it is for the consonants.

The total bandwidth occupancy of the eight control signals is about the same as that of the channel vocoder. If more than four analysis bands are used then the transmission bandwidth can be reduced below that of the channel vocoder.
An alternative configuration is the cascade synthesiser shown in figure 3.16. The control data is the pitch $F_0$; amplitude of voicing $A_v$; three formant frequencies $F_1$, $F_2$, $F_3$ (covering the range approximately 100 to 300 Hz); a single, relatively broad fricative noise resonance $F_n$ (the major resonance in the range 3000-70000 Hz) and the amplitude of noise excitation $A_n$.

All the voiced sounds are produced by the upper resonator string of the circuit, following strictly the cascade approach. The unvoiced sounds are produced by a cascade-parallel connection which introduces zeros as well as poles into the output.

Although the band saving is high, detailed articulation testing of the system showed its performance to be relatively poor. The synthesiser was found to be congenitally incapable of simulating voiced-stops and nasals perhaps because the accuracy required in specifying the frequency of the synthesised zeros was not attained by the system.

A later version of the synthesiser aims to correct some of the shortcomings. It provides for an additional pole-zero pair in the voiced branch and a controllable zero in the unvoiced branch (see figure 3.17). The formant analysis is accomplished by matching the real speech spectrum to a pole-zero model spectrum. There is also more detailed accounting for the excitation characteristics by means of an additional pole-zero pair which contributes significantly to the quality of the synthetic speech.
The vocal transmission for vowel sounds contain only poles and the residues of these poles are functions of the pole frequencies. Therefore, given the formant frequencies, any formant amplitude specification is redundant because the amplitudes are implied by the frequencies. As far as the comparison of the parallel and cascade synthesiser is concerned, the cascade synthesiser provides the correct formant amplitudes automatically from formant frequency data alone. For non-vowel sounds the vocal transmission can have zeros, one or two of which may prove to be perceptually significant. To simulate these factors, the cascade synthesiser requires controllable anti-resonances (ref 3). Again, given the proper pole and zero frequencies, the spectral amplitudes are automatically accounted for. The parallel synthesiser, on the other hand, requires the significant pole frequencies and, ideally, the complex residues of these poles. The residues specify the spectral zeros.

A relevant question about formant synthesis is then "Which is easier to analyse automatically, the frequencies of the spectral zeros or the amplitude and phases of the spectral maxima?" This question is also complicated by the matter of the excitation source. Also what might be asked is "What are the most perceptually salient characteristics?" At present, the ultimate practical choice is still uncertain.

3.3.4 Orthogonal Function Vocoder

One approach in describing a signal with the fewest independent parameters is to approximate the signal by a series of orthogonal
functions. The coefficients of the expression then become the information-bearing quantities. The orthogonal functions chosen for the representation should presumably capitalise upon some known characteristic of the signal.

One orthogonal function description of the short time amplitude spectrum is particularised by Fourier series description as shown in figure 3.18. A short time amplitude spectrum is produced as a time function by scanning at a frequency \(1/P_o\) where \(P_o\) is the pitch period. The frequency \(1/P_o\) would normally range from 25 to 50Hz, the spectral description of \(s(t)\) is transmitted over the restricted bandwidth channel. A bandwidth between 75 and 250Hz is reported to be adequate. Excitation information (i.e. pitch and voiced-unvoiced indications) must also be transmitted. As in the conventional vocoder, a bandwidth of 25-50Hz is adequate for this data. Synchronising information about the scanning must also be made known to the receiver.

At the receiver, a Fourier series description of the amplitude spectrum is computed, namely

\[
s(t) = \frac{a_0}{2} + \sum_{n=1}^{N} \left[ a_n \cos(n\Omega t) + b_n \sin(n\Omega t) \right]
\]

where the coefficients are

\[
a_n = \frac{2}{T} \int_{0}^{T} s(t) \cos(n\Omega t) \, dt
\]

\[
b_n = \frac{2}{T} \int_{0}^{T} s(t) \sin(n\Omega t) \, dt
\]

and \(\Omega = 2\pi/P_o\)

(3.22)
Practically, the Fourier coefficients are obtained by multiplying \( s(t) \) by the outputs of several harmonic oscillators each synchronised to the scanning frequency \( \Omega \). An \( N=3 \) to 5 is claimed to provide an adequate spectral description (PIROGOV)(ref 102).

The coefficients vary relatively slowly in time and are used to control an electrical network so that its frequency response is approximately the same as the measured spectral envelope of the speech signal. The network is then excited in a manner similar to the conventional vocoder, i.e. by periodic pulses or by noise.

### 3.3.5 Homomorphic Vocoders

Homomorphic filtering (ref 30) is a generic term applying to a class of systems in which a signal is transformed into a form where the principles of linear filtering may be applied.

If it is assumed that for speech the excitation sources and the vocal tract shape are relatively independent, a reasonable model is as shown in figure 3.19. In this discrete time model, samples of the speech wave are assumed to be the output of a time-varying digital filter that approximates to the transmission properties of the vocal tract. Since the vocal tract changes shape rather slowly in continuous speech, it is reasonable to assume that this digital filter has fixed characteristics over a time interval of the order of 10mS. Thus the digital filter may be characterised in each time interval by an impulse response or a frequency response or a set of coefficients for an infinite duration impulse response filter.
Specifically, for voiced sounds (except nasals), the transfer function of the digital filter consists of a vocal tract component:

\[
V(z) = \frac{A}{\prod_{k=1}^{p} (1 - c_k z^{-1})(1 - c_k^* z^{-1})} \tag{3.23}
\]

where the \(c_k\)'s correspond to the natural frequencies of the vocal tract, and an additional component

\[
G(z) = B \prod_{k=1}^{m_1} (1 - a_k z^{-1}) \prod_{k=1}^{m_0} (1 - b_k z^{-1}) \tag{3.24}
\]

This accounts for the fact that the finite-duration glottal pulses are not impulses. Thus in figure 3.19 the system function of the digital filter is

\[
H_v(z) = G(z).V(z) \tag{3.25}
\]

This filter is excited by a train of impulses \(p_i\), in which the spacing between the impulses corresponds to the fundamental (or pitch) period of the voice. Unvoiced speech has zeros as well as poles and in this instance a reasonable model is:

\[
H_u(z) = \frac{A \prod_{k=1}^{m} (1 - \alpha_k z^{-1})(1 - \alpha_k^* z^{-1})}{\prod_{k=1}^{p} (1 - c_k z^{-1})(1 - c_k^* z^{-1})} \tag{3.26}
\]

where \(|c_k| < 1\).
In this case the system is excited by a random-noise generator \( r_i \). In both voiced and unvoiced cases, an amplitude control regulates the intensity of the output of the digital filter.

Homomorphic deconvolution can be applied to the estimation of the parameters of the speech model if we assume that the model is valid over a short time interval. Thus a short segment of voiced speech can be thought of as a convolution

\[
\hat{x}_i = p_i \ast g_i \ast v_i \quad 0 \leq i \leq L-1 \tag{3.27}
\]

In order to minimise the effect of the discontinuities at the beginning and end of the interval, a data window \( w_i \) multiplies \( \hat{x}_i \) so that the input to the homomorphic system is

\[
\tilde{x}_i = \hat{x}_i \cdot w_i \tag{3.28}
\]

If \( w_i \) varies slowly with respect to the term \( g_i \ast v_i \), then

\[
\tilde{x}_i = p_{w_i} + \left[ g_i \ast v_i \right] \tag{3.29}
\]

where

\[
p_{w_i} = w_i \cdot p_i
\]

The 'complex cepstrum' of \( x_i \) is obtained by taking the inverse transform of \( \log_e (X'(z)) \); the real cepstrum of \( x_i \) is obtained by taking the inverse discrete Fourier transform of \( \log_e \{|X'(j\omega)|\} \). It can be shown (ref 30) that if \( p_i \) is periodic with period \( P_0 \) then the periodicity manifests itself in the cepstrum in terms of
impulses spaced at intervals of $P_0$ samples. For voiced speech, the
sequence $p_i$ is quasi-periodic hence the first impulse in the
cepstrum is the most prominent. For unvoiced speech, $p_i$ is not
periodic, thus the cepstrum contains no impulses.

In general, the components of the cepstrum due to $v_i$ and $g_i$ decay
rather rapidly, so that for reasonably large values of $P_0$, the
vocal tract and glottal pulse contributions do not overlap the
impulses due to $p_i$. In other words, in the complex logarithm, the
vocal tract and glottal components are slowly varying and the pitch
components are rapidly varying.

If the components of the speech waveform are to be separated then
the complex logarithm should be lowpass filtered to obtain $v_i \ast g_i$,
and highpass filtered to obtain $p_i$. The previous discussion has
shown that homomorphic deconvolution can be successfully applied in
separating the components of a speech waveform which is the
principle of the Homomorphic Vocoder.

The analyser and synthesiser operations for a complete homomorphic
vocoder are shown in figure 3.20. Figure 3.20(a) illustrates the
analysis. At successive intervals (typically every 20mS), the input
speech signal is multiplied by a data window (e.g. a 40mS Hamming*

\[
* \text{The Hamming Window is defined as } w(t) = \left\{ \begin{array}{ll}
0.54 + \cos\left(\frac{2\pi t}{\tau}\right) \\
\end{array} \right.
\]

for \(-\frac{\tau}{2} \leq t \leq \frac{\tau}{2}\)

where $\tau$ is the window duration.
window in this case) and the short time Fourier transform (section 3.3.11) is computed. For each analysis interval the logarithm of the spectral magnitude is taken to produce the log-spectrum $|X'(j\omega)|$. The inverse Fourier transform produces $c(t)$ i.e. the cepstrum of $x(t)$. The final step in the analysis is to derive an equivalent minimum phase description of the vocal tract transmission by truncating and saving the positive low-time part of the cepstrum. (ref 30) This is accomplished by multiplication with the time window $h(t)$. The result is $c'(t)$ which together with the excitation information (the high-time part of $c(t)$ indicating the pitch and voiced/unvoiced decision) constitutes the transmission parameters. The transform of $c(t)$ has a spectral magnitude illustrated by the dashed curve in $\hat{X}(j\omega)$.

Synthesis is accomplished from $c'(t)$ and the excitation information as shown in figure 3.20(b). Periodic pulses, generated at the analysed pitch are used for synthesis of voiced sounds and uniformly spaced pulses of random polarity are used for unvoiced sounds. The transmitted $c'(t)$ is Fourier transformed, exponentiated (to undo the log-taking of the analysis), and inverse transformed to yield a minimum phase (ref 30) approximation to the vocal tract impulse response. This impulse response is convolved with the excitation pulses to produce the output signal $\hat{x}(t)$. 

The transmission rate required for this system is about 7800 bits/sec.

3.3.6 Maximum Likelihood Vocoder

The maximum likelihood method (ref 31) attempts to combine the advantages of time-domain processing and formant representation of the spectrum.

An all-pole model of the power spectrum of the speech signal is assumed. The zeros are omitted because of their lesser importance to perception and because their effects can be represented to any accuracy by a suitable number of poles. The synthesiser includes a recursive digital filter as shown in figure 3.21 whose transfer function is

\[ T(z) = \frac{1}{1 + H(z)} \]  

(3.30)

where \[ H(z) = a_1 z^{-1} + a_2 z^{-2} + \ldots + a_p z^{-p} \]

The complex roots of the denominator polynomial are the complex formants plus the poles which represent the zeros that are used approximate the speech signal. The coefficients \( a_i \) of the denominator polynomial are obtained from time domain calculations on samples of a short segment of the speech waveform; namely \( x_1, x_2, x_3, \ldots, x_N \), where \( N >> p \). Under the assumption that the waveform samples \( x_i \) are samples of a random Gaussian process a maximum likelihood estimate is obtained for the \( a_i \)'s. This estimate corresponds to minimisation
of a further function of the logarithmic difference between the power spectrum of the filter \(|T(z)|^2\) and the short-time power spectrum of the signal sample

\[
X(j\omega) = \frac{1}{2\pi N} \left| \sum_{i=1}^{N} x_i \exp(-j\omega iT) \right|^2
\]

(3.31)

where \(T\) is the sampling interval.

This minimisation results in a fit which is more sensitive at the spectral peaks than in the valleys between the formants. Perceptually this is an important feature of the method. The fit of the all-pole model to the envelope of the speech spectrum is illustrated in figure 3.22.

The maximum likelihood estimate of the filter coefficients is obtained from the short-time correlation function:

\[
\psi_j = \frac{1}{N} \sum_{i=1}^{N-j} x_i x_{i+j}, \quad j = 0, 1, \ldots, N-1
\]

(3.32)

by solving the sets of equations

\[
\sum_{i=1}^{p} \psi_{|i-j|} a_i = -\psi_j, \quad j = 1, 2, \ldots, p
\]

(3.33)
The maximum likelihood estimate also produces the amplitude scale factor for matching the speech signal power spectrum, namely

$$A^2 = \frac{p}{j=-p} A_j \psi_j$$

(3.34)

where

$$A_i = \frac{p}{j=0} a_j a_{j+1}$$

$$a_0 = 1, a_k = 0 (k > p)$$

As shown in figure 3.21, excitation of the synthesiser follows the vocoder convention and uses a pulse generator and a noise generator of the same average power. Extraction of the pitch period $P_0$ (to drive the pulse generator) is accomplished by a modified correlation method which has advantages similar to the cepstrum method, but relies strictly upon time domain techniques and does not require transformation to the frequency domain. A voicing amplitude signal, $V$, is also derived by the pitch extractor. The voiced and voiceless excitations are mixed according to the amplitude of the voicing signal $V$. The unvoiced (noise) excitation level is given by $U = 1 - V^2$. The mixing ratio therefore contains constant average excitation power. Overall control of the mixed excitation by the amplitude signal $A$, completes the synthesis.

Typical parameters for the analysis and synthesis are; sampling rate of input speech, $1/T = 8$ kHz; number of poles, $p=10$; and the number of analysing samples, $N_s = 240$ (i.e. 30mS duration). For transmission purposes, the control parameters are quantized to 9 bits for each of the 10 $a_i$'s, and 6 bits for each of the three
excitation signals \((A, V\) and \(P_0\)). Sampling these quantized parameters at 50 sec\(^{-1}\) yields a 5400 bit/sec encoding of the signal for digital transmission. The technique is demonstrated to be substantially better than digitized channel vocoders. \(\text{(ref 31)}\)

### 3.3.7 Linear Prediction Vocoder

The analysis employed in the linear prediction coding (LPC) vocoder is another time domain technique. Being a time domain technique this again avoids the formant location difficulties of the frequency domain formant analysis methods, where formants seem to disappear during certain sounds or seem to increase their number during others. The LPC method again utilises an all-pole recursive digital filter excited either by a pitch-modulated pulse generator or a noise generator to synthesise the signal as with the maximum likelihood vocoder. The filter coefficients in this case represent an optimum linear prediction of the signal. The coefficients are determined by minimising the mean square error between samples of the input signal and signal values estimated from a weighted linear sum of past values of the signal. That is for every sample of the input signal, \(x_i\), an estimate \(\hat{x}_i\) is formed such that

\[
\hat{x}_i = \sum_{k=1}^{P} a_k x_{i-k}
\]

The filter coefficients, \(a_k\), are determined by minimising \((x_i - \hat{x}_i)^2\) over an analysis interval that is typically a pitch period, but
which may be as small as 3mS for p=12 and a sampling rate of 10kHz. The $a_k$'s are given as a solution of the matrix equation

$$\phi A = \Psi$$  \hspace{1cm} (3.35)

where $A$ is a $p$-dimensional vector whose $k^{th}$ component is $a_k$, $\phi$ is a (pxp) covariance matrix with term $\phi_{ij}$ given by

$$\phi_{ij} = \sum_{n=1}^{p} x_{n-j} x_{n-i} \hspace{1cm} (j=1,2,\ldots,p)$$  \hspace{1cm} (3.36)

and $\Psi$ is a $p$-dimensional vector with the $j^{th}$ component $\psi_j = \phi_{j0}$. Since the matrix $\phi$ is symmetric and positive definite, equation (3.35) can be solved without matrix inversion. These relations are similar to those obtained from the Maximum Likelihood method except for the difference in the matrix $\phi$. The two solutions approach each other for the condition $N >> p$.

Synthesis is accomplished as shown in figure 3.23. Excitation either by pitch-modulated pulses or by random noise is supplied to a recursive filter formed from the linear predictor. The amplitude, $A$, of the excitation is derived from the rms value of the input speech wave. The filter transmission function is

$$T(z) = \frac{1}{1 - H(z)}$$  \hspace{1cm} (3.37)

where

$$H(z) = \sum_{k=1}^{p} a_k z^{-k}$$
which except for the sign convention, is the same as the Maximum Likelihood method. The filter coefficients $a_k$ account both for the filtering of the vocal tract and the spectral properties of the excitation source. If $e_i$ is the $i^{th}$ sample of the excitation, then the corresponding output sample of the synthesiser is

$$x'_i = e_i + \sum_{k=1}^{p} a_k x'_{i-k}$$

(3.38)

where the primes distinguish the synthesised samples from the original speech samples. The complex roots of $[1 - H(z)]$ in equation (3.37) therefore include the bandwidths and frequencies of the speech formants. The filter coefficient data can be transmitted directly as the values of the $a_k$, or in terms of the roots of $[1 - H(z)]$. The latter requires a root finding calculation.

A method of pitch extraction uses the error $(x_i - \hat{x}_i)$ and a peak-picking algorithm to determine the pitch period. Good quality synthesis at a digital bit rate as low as 3600 bits/sec have been used with $p=12$.

The LPC vocoder and the maximum likelihood vocoder have been found useful for automatic formant extraction. (ref 32)

3.3.8 Articulatory Vocoder

Another approach to the general vocoder problem is to code speech in terms or articulatory parameters; such a description has the advantage of duplication, on a direct basis, of the physiological
constraints that exist in the human vocal tract. Non-discontinuous signals that describe the vocal transmission would then produce all sounds, consonants and vowels.

The notion is to transmit a set of data which describes the tract configuration and its excitation as functions of time. The synthesiser can be a controllable vocal tract model. The success of this system will depend largely upon the precision with which articulatory data can be obtained automatically from the acoustic signal.

The three following vocoder systems now discussed avoid the pitch tracking and the voiced/unvoiced decision problem.

3.3.9 Frequency Dividing Vocoder

Frequency division is a well-known process for reducing the bandwidth of signals whose spectral widths are determined primarily by large index frequency modulation. While speech is not such a signal, sub-bands of it (e.g. formant bands or individual voice harmonics (ref 3)) have similarities to large-index frequency modulation. Frequency division by factors of two or three are possible before intelligibility substantially deteriorates.

Frequency division generally implies possibilities for frequency multiplication and similarly, spectral division-multiplication processes suggest possibilities for compression and expansion of the signal's time scale. Reproduction of a divided signal at a rate
proportionately faster restores the frequency components to the original values and compresses the time scale by a factor equal to the frequency divisor.

Various methods - including electrical, mechanical, optical and digital - have been used to accomplish division and multiplication. A method of rejecting portions of speech segments (proposed by GABOR in 1940[ref 33]) was achieved by implementing a magnetic tape head assembly in a rotating wheel and scanning a magnetic tape. This was later developed by Fairbanks et al[ref 34] which employed multiple pick-up heads in a rotating drum. The differential tape-to-drum speeds could be varied independently to attain any degree of speech compression or expansion. Variable speech control (VSC)[ref 35] is based upon deletion of audio signals and stretching remaining segments to reproduce the sound. It is claimed that 20-40mS per phoneme can be deleted without the loss of auditory continuity. Alternatively[ref 36] it is possible to perform time compression by the elimination of the pauses between words that exceed a preset time interval. However, the natural pause-to-speech sound relationship at the juncture of phrases and even intermittent hesitation between words should be maintained for maximum listening comprehension. This method can also be achieved with buffer stores with variable read-in and fixed read-out rates to vary the word rate and keep the pitch constant.[ref 36]
3.3.9.1 Vobanc

One frequency division method for bandwidth compression is the 
vobanc (ref 37). Although constructed practically using heterodyne 
techniques, the principle involved is as shown in figure 3.24. The 
speech band 200 to 3200kHz is separated into three contiguous band 
pass channels $A_1$, $A_2$ and $A_3$.

Each channel is about 1kHz wide and normally covers a range of a 
speech formant. Using a regenerative modulator, the signal in each 
band is divided by two and limited to half the original frequency 
range by bandpass filters $B_1$, $B_2$ and $B_3$. The added outputs of 
the filters yield a transmission signal which is confined to about 
half the original bandwidth.

At the receiver, the signal is again filtered into three bands, 
$B_1$, $B_2$ and $B_3$. The bands are restored by frequency doubling 
and are combined to provide the output signal. In consonant 
articulation tests, the Vobanc consonant articulation was 
approximately 80 per cent, using 48 listeners and 10 talkers. In 
the same test, an otherwise unprocessed signal band limited 200 to 
1700Hz scored a consonant intelligibility of 66 per cent.

3.3.10 Analytic Rooter

Another technique for frequency division of formant bands of speech 
is called analytic rooting (refs 38,39) where the processing is
done in terms of the analytic signal. This approach avoids characteristic degradations that frequency division methods such as those that the Vobanc introduce.

The analytic signal $\sigma(t)$ of a real, bandlimited signal $x(t)$ is defined as

$$\sigma(t) = x(t) + j \bar{x}(t)$$  \hspace{1cm} (3.39)

where $\bar{x}(t)$ is the Hilbert transform of $x(t)$.

The Hilbert transform is defined as

$$\bar{x}(t) = \frac{1}{\pi t} \ast x(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau$$  \hspace{1cm} (3.40)

The Hilbert transformer, shown in figure 3.25, is also a quadrature phase shifter.

In polar form the analytic signal is

$$\sigma(t) = a(t). \exp\{j \phi(t)\}$$  \hspace{1cm} (3.41)

where

$$a(t) = \left[ x^2(t) + \bar{x}^2(t) \right]^{1/2} \hspace{1cm} \text{(The Hilbert envelope)}$$

and

$$\phi(t) = \tan^{-1} \left\{ \frac{\bar{x}(t)/x(t)}{1} \right\}$$
It follows that
\[ x(t) = a(t) \cos[\phi(t)] \]
and
\[ \bar{x}(t) = a(t) \sin[\phi(t)] \tag{3.42} \]

A real signal \( x_{1/n}(t) \) corresponding to the \( n \)th root of the analytic signal can be defined as
\[
x_{1/n}(t) = \text{Re} \left\{ \sigma(t) \right\}^{1/n}
\]
\[
= \text{Re} \left\{ x(t) + j \bar{x}(t) \right\}^{1/n}
\]
\[
= \left[ a(t) \right]^{1/n} \cdot \cos \left[ \phi(t)/n \right] \tag{3.43}
\]

The instantaneous frequency \( f(t) \) is
\[
f(t) = \frac{1}{2\pi} \frac{d}{dt} [\phi(t)]
\]
\[
\therefore \frac{f(t)}{n} = \frac{1}{2\pi} \frac{d}{dt} [\phi(t)/n] \tag{3.44}
\]

The analytic signal rooting therefore implies that the frequency bandwidth of \( x(t) \) is compressed into \( 1/n \) and that it is possible to deal separately with the amplitude and phase components (the frequency components) of the signals. For those cases where the perceived pitch is determined by the envelope of the signal waveform, this process can leave the pitch unaltered. This method is therefore attractive for restoring speech distorted by a helium atmosphere, such as that breathed by a deep-sea diver.

For the case of \( n=2 \)
\[
x_{1/2}(t) = \left[ a(t) \right]^{1/2} \cdot \cos \left[ \frac{1}{2} \phi(t) \right]
\]
\[
= [a(t)]^{1/2} \cdot \left[ \frac{1}{2} (1 + \cos \phi(t)) \right]^{1/2}
\]

Since \( a(t) \cos \phi(t) = x(t) \)

\[
x_{1/2}(t) = \left[ \frac{1}{2} \right]^{1/2} \left[ a(t) + x(t) \right]^{1/2} \tag{3.45}
\]

Similarly, it can be shown that the Hilbert transform \( \bar{x}_{1/2}(t) \) of \( x_{1/2}(t) \) is

\[
\bar{x}_{1/2}(t) = \left[ \frac{1}{2} \right]^{1/2} \left[ a(t) - x(t) \right]^{1/2} \tag{3.46}
\]

Equation (3.46) also follows from (3.45) by the observation that multiplication of \( x(t) \) by \(-1\) is equivalent to a phase shift of \( \pi \) radians, which corresponds to a phase shift of \( \pi/2 \) in \( x_{1/2}(t) \) i.e. a Hilbert transformation.

The proper sign of equation (3.45) can be recovered by changing the sign of the square root (3.45) every time the phase \( \phi(t) \) of the original signal \( x(t) \) goes through \( 2\pi \) radians (or an integer multiple of \( 2\pi \)). According to (3.42) this is the case when \( \bar{x}(t)=0 \), while \( x(t)>0 \).

The inverse operation of analytic-signal rooting is given by

\[
x_n(t) = \text{Re} \left\{ \left[ x(t) + j \bar{x}(t) \right]^n \right\}
= \left[ a(t) \right]^n \cdot \cos \left[ n \phi(t) \right] \tag{3.47}
\]

For \( n=2 \)

\[
x_2(t) = \text{Re} \left\{ x^2(t) + 2j x(t) \bar{x}(t) - \bar{x}^2(t) \right\}
= x^2(t) - \bar{x}^2(t) \tag{3.48}
\]
If the inverse process is applied to \( x_{1/2}(t) \), the original signal \( x(t) \) should be recovered. This can be verified by substituting \( x_{1/2}(t) \) and \( \bar{x}_{1/2}(t) \) from (3.45) and (3.46) into (3.48).

\[
x_2(t) = \frac{1}{2} \left\{ (a(t) + x(t)) - (a(t) - x(t)) \right\}
\]

or \( x_2(t) = x(t) \) \( (3.49) \)

The Hilbert transform of the original signal can be recovered by multiplying \( x_{1/2}(t) \) and \( \bar{x}_{1/2}(t) \)

\[
2x_{1/2}(t).\bar{x}_{1/2}(t) = \left\{ (a(t) + x(t))(a(t) - x(t)) \right\}^{1/2}
= \left\{ a^2(t) - x^2(t) \right\}^{1/2}
= \bar{x}(t)
\]

\( (3.50) \)

For a signal whose bandwidth is narrow compared to its centre frequency, the original signal can be approximately recovered by squaring \( x_{1/2}(t) \) and subsequently bandpass filtering.

\[
2x^2_{1/2}(t) = a(t) + x(t)
\]

\( (3.51) \)

If the spectrum of \( a(t) \) (the envelope function) does not overlap that of \( x(t) \), then \( x(t) \) can be recovered by bandpass filtering.

A complete transmission system based upon the foregoing principles is shown in figure 3.26. The speech spectrum is first divided into four contiguous passbands, each nominally containing no more than one formant. Each bandpass signal is then analytically rooted, band-limited, and recovered.
The Hilbert transform of each bandpass signal is formed by a transversal filter HT1 of impulse response \( h(t) = 1/\pi t \). Since the Hilbert transform filter ideally has a response which is neither time limited nor band limited, an approximation is made to the transform which is valid over the frequency range of interest and which is truncated in time.

In a parallel path, the bandpass signal \( x(t) \) is delayed by an amount \( DEL1 \) equal to half the duration of the impulse response of the Hilbert filter. The two signals are squared and summed to give \( (x^2 + \bar{x}^2) \) at the output of ADD1. The square root of this result yields \( a(t) \), and the addition of the delayed \( x(t) \) in ADD2 gives \([a(t)+x(t)]\). Attenuating the signal by 1/2 and the subsequent square rooting forms \( x_{1/2}(t) \).

Selection of the sign of \( x_{1/2}(t) \) is accomplished by the following logical decisions in the SWITCH. The algebraic sign of \( x_{1/2}(t) \) is changed whenever \( \bar{x}(t) \) goes through zero and \( x(t) > 0 \). The signal \( x_{1/2}(t) \) is then applied to BPF 1/2, having its cut off frequencies and hence bandwidth equal to half the values for BPF1.

At the receiver, analytic squaring of this band limited version of \( x_{1/2}(t) \) is accomplished in accordance with (3.48). The Hilbert transform is produced by HT2 which is similar to HT1 except that the duration of the impulse response of the former is twice that of the latter. Subtracting \( \bar{x}^2(t) \) from \( x^2(t) \) recovers an approximation to the original bandpassed signal \( x(t) \).
The simulation operations (ref 38) in all four channels are identical except that the bandpass filters, Hilbert transform filters, and delays are chosen in accordance with the desired passband characteristics. Eighth-order Butterworth filters with cut off frequencies listed in Table (3.2) were used for the bandpass filters. The result is said to be of a respectable quality over a channel band limited to one-half of that of the original signal.

3.3.11 Phase Vocoder

The phase vocoder (ref 40) makes use of the short-time phase derivative spectrum of the signal to accomplish the band saving; the method permitting non-integer divisions as well as integer values. It can be applied either to single voice harmonics or to wider sub-bands which can include single formants. It also permits a flexible means for time compressing or expanding of the speech signal.

If the speech signal \( x(t) \) is passed through a parallel bank of contiguous bandpass filters and then recombined, the signal is not substantially degraded; this is illustrated in figure (3.27). The filters are assumed to have a relatively flat amplitude and linear phase characteristics in their passbands. The output of the \( n \)-th filter is \( x_n(t) \), and the original signal is approximated as

\[
\hat{x}(t) = \sum_{n=1}^{N} x_n(t) \quad (3.52)
\]
Let the impulse response of the n-th filter be

\[ g_n(t) = h(t) \cos \omega_n t \] (3.53)

where the envelope function \( h(t) \) is normally the impulse response of a physically-realisable lowpass prototype filter. The output of the n-th filter is the convolution of \( x(t) \) with \( g_n(t) \)

\[ x_n(t) = \int_{-\infty}^{t} x(\lambda) h(t-\lambda) \cos[\omega_n (t-\lambda)] d\lambda \]

\[ = \text{Re} \left[ \exp(j\omega_n t) \int_{-\infty}^{t} x(\lambda) h(t-\lambda) \exp(-j\omega_n \lambda) d\lambda \right] \] (3.54)

The latter integral is a short-time Fourier transform of the input signal \( x(t) \), evaluated at radian frequency \( \omega_n \). It is the Fourier transform of that part of \( x(t) \) which is 'viewed' through the sliding time aperture \( h(t) \). (ref 41) If the complex value of this transform 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 \( \phi(\omega_n, t) \). Then

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

i.e.

\[ x_n(t) = |X(\omega_n, t)| \cdot \cos[\omega_n t + \phi(\omega_n, t)] \] (3.55)

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 \( \omega_n \).
Experience with channel vocoders shows that magnitude functions 
\[ |X(\omega_n, t)| \] may be bandlimited to around 20 to 30Hz without 
substantial loss of perceptually-significant detail. The phase 
functions \( \phi(\omega_n, t) \), however, are generally not bounded; hence they 
are unsuitable as transmission parameters. Their time 
derivatives \( \dot{\phi}(\omega_n, t) \), on the other hand, are more well behaved, and 
may be band limited and used to advantage in transmission. To 
within an additive constant, the phase functions can be recovered 
from the integrated (accumulated) values of the derivatives. A 
practical approximation to \( x_n(t) \) is therefore

\[
\dot{x}_n(t) = |X(\omega_n, t)| \cos \left[ \omega_n t + \dot{\phi}(\omega_n, t) \right] 
\tag{3.56}
\]

where
\[
\dot{\phi}(\omega_n, t) = \int_0^t \dot{\phi}(\omega_n, t) dt
\]

Reconstruction of the original signal is accomplished by summing the 
outputs of the \( n \) oscillators modulated in phase and amplitude. The 
oscillators are set at the nominal frequencies \( \omega_n \) and they are 
simultaneously phase and amplitude modulated from band limited 
versions of \( \dot{\phi}(\omega_n, t) \) and \( |X(\omega_n, t)| \) which is illustrated in figure 
(3.28).

The conventional channel vocoder separates vocal excitations and 
spectral envelope functions. The spectral envelope functions of the 
conventional vocoder are the same as those described by \( |X(\omega_n, t)| \) in 
the phase vocoder. The excitation information, however, is
contained in a signal which specifies voice pitch and voiced-unvoiced (buzz-hiss) excitation. In the phase vocoder, when the number of channels is reasonably large, information about excitation is conveyed primarily by the \( \dot{\phi}(\omega_n, t) \) signals. Alternatively, one can consider the \(|X(\omega_n, t)|\) signals indicating the spectral shape; the number of contiguous filters effectively quantizing the spectral envelope. The \( \dot{\phi}(\omega_n, t) \) signals, being orthogonal to the \(|X(\omega_n, t)|\) plane, indicate the relative deviation of the actual maxima (pitch harmonic) from the centre frequency, \( \omega_n \) of each respective bandpass filter, i.e. frequency correction.\(^{\text{(ref 2)}}\) If good quality and natural transmission are requisites, the indications are that the \( \dot{\phi}(\omega_n, t) \) signals require about the same channel capacity as the spectrum envelope information.

In the implementation/simulation of the phase vocoder, the amplitude and phase spectrum at the analyser are computed by forming the real and imaginary parts.

\[
X(\omega_n, t) = a(\omega_n, t) - j b(\omega_n, t) \tag{3.57}
\]

where

\[
a(\omega_n, t) = \int_{-\infty}^{t} x(\lambda) h(t-\lambda) \cos \omega_n \lambda \, d\lambda
\]

and

\[
b(\omega_n, t) = \int_{-\infty}^{t} x(\lambda) h(t-\lambda) \sin \omega_n \lambda \, d\lambda
\]

then

\[
|X(\omega_n, t)| = (a^2 + b^2)^{1/2}
\]

and

\[
\phi(\omega_n, t) = \tan^{-1} \left( \frac{-b}{a} \right)
\]

from which

\[
\dot{\phi}(\omega_n, t) = \frac{ab - ba}{a^2 + b^2} \tag{3.58}
\]
Transforming the real and imaginary parts of (3.57) into discrete form for programming yields;

\[ a(\omega_n, mT) = T \sum_{\ell=0}^{m} x(\ell T) \left[ \cos \omega_n \ell T \right] \cdot h(mT - \ell T) \]  \hspace{1cm} (3.59)

\[ b(\omega_n, mT) = T \sum_{\ell=0}^{m} x(\ell T) \left[ \sin \omega_n \ell T \right] \cdot h(mT - \ell T) \]  \hspace{1cm} (3.60)

where \( T \) is the sampling interval. The difference values are computed as:

\[ \Delta a = a\left[ \omega_n, (m+1)T \right] - a\left[ \omega_n, mT \right] \]  \hspace{1cm} (3.61)

\[ \Delta b = b\left[ \omega_n, (m+1)T \right] - b\left[ \omega_n, mT \right] \]  \hspace{1cm} (3.62)

The magnitude function and the phase derivative in discrete form are:

\[ |X\left[ \omega_n, mT \right]| = (a^2 + b^2)^{1/2} \]

\[ \frac{\Delta \phi}{T} \left[ \omega_n, mT \right] = \frac{1}{T} \frac{(b \Delta a - a \Delta b)}{a^2 + b^2} \]  \hspace{1cm} (3.63)

Figure 3.29 shows a block diagram of a single analyser channel which is required for each channel (typically there are about 30 channels). The equivalent pass bands of the analysing 6th order Bessel filters overlap at the -6dB points, and a total spectrum range of 50 to 3050Hz is analysed where \( \omega_n = 2\pi \left( 100n + 50 \right) \) i.e. 100Hz separation.
The synthesis operation for each channel is performed according to:

\[ x_n(mT) = |X(\omega_n,mT)| \cos \left[ \omega_n,mT + T \sum_{k=0}^{m} \frac{\Delta \phi_k}{T} (\omega_n,kT) \right] \]  \hspace{1cm} (3.64)

Adding the outputs of the \( n \) individual channels produces the synthesised speech signal. A frequency divided signal may be synthesised by division of the \([\omega_n t + \int \phi_n dt]\) quantities by some number \( i.e. \) in figure 3.28, divide \( \phi \) by \( q \) and use \( \cos \left[ \frac{1}{q} \omega_n + \frac{\phi}{q} \right] \) for the synthesiser.

This frequency-divided synthetic signal may be restored to its original spectral position by a time speed-up of \( q \).

Time scale expansion of the synthesised signal is likewise possible by the frequency multiplication \( q[\omega_n t + \int \phi_n dt] \) then by recording the frequency multiplied signal and playing it \( q \)-times slower.

The frequency division and multiplication factors can be non-integers and can be varied with time therefore the phase vocoder provides an attractive tool for studying non-uniform alterations of the time scale.

A number of multiplexing methods may be used for transmission; conventional space, frequency and time-division methods are obvious techniques. A "self-multiplexing" method is also possible in which, say, a 2 to 1 frequency divided synthetic signal is transmitted over an analogue channel of half of the original signal bandwidth.
Re-analysis, frequency expansion and synthesis at the receiver recovers the signal. The greatest number of \( q \) by which the \( \omega_n \) and \( \phi_n \)'s may be divided is determined by how distinct the side bands about each \( \omega_n/q \) remain, and by how well each \( \phi_n/q \) and \( |X_n| \) can be retrieved from them. Practically, the greatest number appears to be about 2 or 3 if transmission of acceptable quality is to be realised. In one digital implementation, the phase and amplitude functions were sampled, quantized and framed from digital transmission at rates of 9.6kb/s and 7.2kb/s. These transmission rates were compared in listening tests to the same signal coded by log-PCM. The results showed the digital phase vocoder to provide a signal quality comparable to log-PCM at bit rates two or three times higher.\(^\text{(ref 42)}\) The quality of the resulting signal considerably surpasses that usually associated with conventional channel vocoders.

The implementation of the digital phase vocoder can be simplified computationally by the use of the Fast Fourier Transform (FFT) algorithm both in the analysis and synthesis procedures. When this is operated as an identity system (i.e. no compression or expansion), the synthesised output differs in no perceptual or measurable way from the input speech.\(^\text{(ref 43)}\) Other computer implementations of the phase vocoder have also been recently devised.\(^\text{(refs 44-46)}\) These consider sub-band coding (see section 3.4.1.2) for the bands 200-400 and 400-800Hz and dividing the remaining frequency range 800-3200Hz into ten 1/6-octave bands in accordance with a 'perception-specified' criterion (based upon critical bands implied by the ear) rather than a speech specific
criterion. The phase vocoder operations are applied to each of the sub-bands to yield the amplitude $|X_n|$ and phase derivative $\dot{\phi}_n$ information which are then coded using 3 or 4-bit APOM or ADPCM in various combinations to obtain transmission bit rates of between 16-20 kblts/s.

It is stated that the method can give reasonably acceptable quality but at lower bit rates than 16-20 kb/S into the data speed range intelligibility remains useful but talker recognition and quality are significantly degraded. As before, the derived parameters $|X_n|$ and $\dot{\phi}_n$ offer a means of time-scale modification without attendant frequency shifts.

3.4 Intermediate System

There are many ways of combining some of the signal description possibilities of waveform coders with some of the signal redundancy exploitation of vocoders. The resultant intermediate systems normally give better speech reproduction in the 4-16 kb/s range than is possible with either of the other two classes of system at these digit rates. Their complexity is always greater than for simple waveform coders and some of the higher performance systems may be more complicated than vocoders. Below are some examples of intermediate systems.
3.4.1 Frequency-Domain Coding

3.4.1.1 Voice-Excited Vocoder (VEV)

One useful approach to gaining some of the advantage of waveform coding and vocoders is to use vocoder techniques for the perceptually less critical high frequency regions, and to use waveform coding for the low frequency regions up to, say, 800Hz. (refs 47,48) This also avoids the difficulties inherent in automatic analysis of the excitation data. At the receiving end, the unprocessed baseband is put through a non-linear distortion process to spectrally flatten and broaden it. It is then used as the source of excitation for regular vocoder channels covering the frequency range above the baseband. A block diagram of the arrangement is shown in figure 3.30.

The flattened excitation band reflects the spectral line structure of the quasi-periodic voiced sounds and the continuous spectral character of the unvoiced sounds. Because it is derived from a real speech band, it inherently preserves the voiced-unvoiced and pitch information.

In one implementation of the device the baseband is taken as 250 to 940Hz. The frequency range 940 to 3650Hz above the baseband is covered by 17 vocoder channels. The first 14 of these channels have analysing bandwidths of 150Hz, and the upper three are slightly wider. The total transmission band occupancy is 1000 to 1200Hz, yielding a bandwidth compression of about three-to-one. One method
of spectral flattening is shown in figure 3.31. The transmitted baseband is rectified and applied to the bandpass filters of the vocoder synthesizer. The filter outputs are peak-clipped to remove amplitude fluctuations. They are then applied as inputs to amplitude modulators which are controlled by the vocoder channel signals.

Intelligibility and speech quality tests, using speech from a carbon microphone, were carried out to compare the voice-excited vocoder to telephone handset speech bandlimited to the same frequency range (ref 48). The results show the voiced-excited system to be better than the spectrum channel vocoder and to approach the quality of conventional voice circuits. Its application, as with similar methods, depends upon desired trade-offs between costs of terminal equipment, amount of band-saving and signal quality.

3.4.1.2 Sub-Band Coding (SBC) (refs 49,50)

This technique offers attractive possibilities for coding speech at bit rates in the range 7.2 to 16kb/s. In the sub-band coder the speech band is divided into typically four to eight sub-bands by a bank of band-pass filters. Table (3.3) shows 4 such sub-bands which have equal contribution to the Articulation Index (AI). (ref 3) Each sub-band is, in effect, low pass translated to zero frequency by a modulation process equivalent to single side-band modulation. It is then sampled (or resampled) at its Nyquist rate (twice the width of the band) and digitally encoded with APCM (section 3.2.1.2) with Jayant's quantization strategy. In this process, each sub-band
can be encoded according to perceptual criteria that are specific to that band. On reconstruction, the sub-band signals are decoded and modulated back to their original locations. They are then summed to give a close replica of the original speech signal.

<table>
<thead>
<tr>
<th>Sub-band Number</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>

Table 3.3  Partitioning of the 200-3200Hz Frequency range into sub bands that give an equal contribution to the Articulation index (ref 3).

Encoding in sub-bands offers several advantages. Quantization noise can be contained in bands to prevent masking of one frequency range by quantizing noise in another frequency range. Separate adaptive quantizer step sizes can be used in each band. Therefore, bands with lower signal energy will have lower quantizer step sizes and contribute less quantization noise. By appropriately allocating the bits in different bands, the shape of quantization noise can be controlled in frequency. In the lower frequency bands, where pitch and formant structure must be accurately preserved, a large number of bits/sample can be used; whereas in upper frequency bands, where fricative and noise like sounds occur in speech, fewer bits/sample can be used.
Figure 3.32 illustrates a basic block diagram of the sub-band coder. Essentially, it consists of a bank of bandpass filters, APCM encoders and a multiplexer. The modulation is obtained essentially "for free" by using the technique of integer-band sampling. The technique is illustrated in figure 3.33. The signal sub-bands $x_n(t)$ are chosen to have a lower cut-off frequency of $mf_n$ and an upper cut-off frequency of $(m+1)f_n$ where $m$ is an integer and $f_n$ is the bandwidth of the $n$-th band. This bandpass signal is sampled at $2f_n$ to produce a sampled spectrum shown in figure 3.33 (for $m=2$). The signal may be then decimated to alias the band down to baseband prior to encoding. After decoding, the band is interpolated to translate it effectively to its original frequency range. The received signal is recovered by decoding and bandpassing to the original signal band. A slight disadvantage is that integer band restrictions prevent the choice of bands strictly on the basis of equal contribution to the Articulation Index (AI). However, little loss in performance is reported (ref 49).

The most complex part of the coder is the filter bank. However, with newer filter technologies such as CCD filters and digital filters, this complexity is rapidly being reduced to a chip level. Also the design technique of quadrature mirror filters (QMF) affords distinct advantages in the digital implementation of this coder; the principle of quadrature mirror filtering is briefly explained as follows (refs 41, 50, 51). The sub-band coder first must divide the input speech spectrum into four bands (usually non-contiguous) as shown in figure 3.32. In order to preserve an even spectral
coverage, i.e. minimal overlap of the filter transition regions, the filter response skirts must be comparatively steep. If the filters are implemented by IIR filters then the phase response will be highly non-linear at the filter transition regions.

This can be overcome by employing a linear phase response obtained by using a FIR filter. However, at the lower part of the speech spectrum, the sub-bands are comparatively narrow (i.e. 100-200Hz) which can only be provided by a long non-recursive filter requiring extensive computation.

The quadrature mirror approach can be developed from a two-band filter bank as shown in figure 3.34(a). The analysis low pass filter \( h(i) \) can be converted into a high pass filter by inverting the sign of alternate samples of the impulse response \( h(i) \) to give \( \hat{h}(i) \). This is also achieved by inverting the sign of every alternate sample of the input signal \( x_i \). This is in fact equivalent to a digital modulation giving a resulting spectral shift by one-half the Nyquist frequency; a spectral interpretation of this process is shown in figure 3.34(b). A careful analysis of the quadrature filter bank reveals that the frequency aliasing (i.e. the leakage illustrated by the shaded region in figure 3.34(b)) cancels to below the level of quantization noise that would normally be expected from the APCM coders. This is due to the fact that \( \hat{H}(j\omega) \) is the negative spectrum of \( H(j\omega) \) such that phase cancellation occurs where these spectra overlap. Because of this cancellation of aliasing terms, the designer of the sub-band coder is no longer obliged to include analysis filters with a steep roll-off in the transition regions. This greatly eases the computation load.
The perceptual advantages of sub-band coding are well put in evidence by the data of figure 3.35 (ref 52). Figure 3.35(a) shows the relative preference of a 16 kbits/s sub-band coder versus that of an ADPCM coder at various ADPCM coder bit rates (a 50-50 per cent preference implies that the two qualities are almost equal). Figure 3.35(b) shows similar results for a 9.6 kbits/s sub-band coder compared against an ADM coder at various ADM bit rates. In comparison with ADM and ADPCM at transmission bit rates in the order of 16 kbits/s one sees that the effect of SBC is worth the equivalent of about 10 kbits/s of transmission rate.

3.4.1.3 Adaptive Transform Coding (ATC) (ref 53)

In transform coding (TC) systems, each block of speech samples is transformed into a set of transform coefficients; these coefficients are then quantized independently and transmitted. An inverse transform is taken at the receiver to obtain the corresponding block of reconstructed speech samples.

A transform coding system is shown in figure 3.36 and operates as follows: A block of N successive input samples \( x_i, i = 1, 2, \ldots \), \( N \) is arranged into the vector \( x \); this vector \( x \) is linearly transformed using matrix \( A \) such that

\[
\mathbf{y} = \mathbf{A} \cdot \mathbf{x}
\]  

(3.65)

The elements of \( y \) are the transform coefficients of the coding scheme; each of these elements is independently quantized, thus
leading to the vector \( y \). This vector of quantized transform coefficients is transmitted to the receiver and transformed using the inverse matrix \( A^{-1} \):

\[
\hat{x} = A^{-1} \cdot \hat{y}
\]  

(3.66)

The elements of \( \hat{x} \) are the reconstructed output samples, and when successive segments are concatenated, they represent the speech signal.

a) The Block Transformation

The class of block-transformations of interest for speech processing are time-to-frequency transformations. Since a primary goal is to generate the least audible coding noise, it is natural to control the quantization noise by controlling its characteristics in the frequency domain. Also, speech production can be modelled on a short-time basis.

A particular time-to-frequency transform, referred to as the Discrete Cosine Transform (DCT), has been found to be well suited for speech coding. One of the attractions of the DCT is that, in a long-time sense, it happens to be a good fit to the optimally orthogonalizing Karhunen-Loève Transform (KLT) for speech waveforms (ref 53). The even symmetry inherent in the structure of the DCT framework also helps to minimize end effects. The DCT of an \( N \)-point sequence \( x_i \) is defined as
where \( g(0) = 1 \) and \( g(k) = \sqrt{2}, \) \( k = 1, ..., N-1. \) The inverse DCT (IDCT) is defined as:

\[
x_i = \frac{1}{N} \sum_{k=0}^{N-1} X_c(k) g(k) \cos \left( \frac{2(i+1)k\pi}{2N} \right)
\]  

(3.68)

Fast algorithms have been defined for implementing the DCT with great computational efficiency.

The choice of analysis window for the DCT is important in controlling the block boundary effects. Trapezoidal shaped windows have been found very useful for low bit rate transform coding of speech. By allowing a small overlap between the successive blocks, such that the sum of the overlapped windows is always unity, an additional lowering of the number of bits is available for encoding each block.

b) Quantization Strategy

The transform coefficients are usually quantized individually using a uniform step size \( \Delta_k \) and a number of levels \( 2^{b_k}. \) The choice
of $\Delta_k$ and the number of bits $b_k$ for a given transform coefficient, $k$, is of fundamental importance. Ideally, $\Delta_k$ must be just large enough so that overloading of the quantizer does not occur. If this is assumed, the number of the quantization levels as a function of coefficient number and the amplitude spectrum determine the coarseness of the quantization and thus the spectral characteristics of the quantization noise.

If the transform coefficients were stationary Gaussian variables with variances $\sigma_k^2$, $k = 0, 1 \ldots N-1$, and if a flat noise spectral distribution is desired, then the optimal bit assignment for the $k^{th}$ coefficient is:

$$b_k = \delta + \frac{1}{2} \log_2 \left( \frac{\sigma_k^2}{D^*} \right)$$

$$k = 0, 1, \ldots, N-1$$

(3.69)

where $\delta$ is a correction term that reflects the performance of practical quantizers, and $D^*$ denotes the noise power.

$$D^* = \frac{1}{N} \sum_{k=0}^{N-1} E_k^2$$

(3.70)

where $E_k^2$ is the noise power incurred in quantizing the $k$-th transform coefficient. These bit assignments have to satisfy the constraint:

$$B = \sum_{k=1}^{N} b_k$$

(3.71)
where \( B \) is the number of bits/block available for transmission over the binary channel.

An interpretation of this bit assignment rule can be seen in figure 3.37 for given positions of the horizontal dashed thresholds, i.e. for given values of spectral levels \( L_m \), an assignment is made so that every coefficient with variance below \( L_0 \) will be assigned zero bits (i.e. it will not be transmitted), those with variances between \( L_0 \) and \( L_1 \) will be assigned 1 bit, and so on. If the resulting total number of bits assigned in this way is less than the total number of bits available for transmission, \( B \), the values of \( L_m \) are increased. This process continues until equation (3.71) is satisfied. Given the number of levels per quantizer and the variance of the transform coefficients, the choice of the optimal step sizes for the uniform quantizers depends only on the PDF of the transform coefficients (a Gaussian PDF is typically assumed).

c) **Adaptation Strategy**

The expected spectral levels \( \hat{\sigma}_k \) of the transform coefficients for any particular speech block are not known \( \text{a priori} \) and therefore must be estimated. This information, which reflects the dynamical properties of speech in the transform domain, is commonly referred to as "side-information".

Two basic adaptation techniques for ATC of speech have been proposed. In one, illustrated in figure 3.38, the side information consists of a small number of samples computed by averaging
(smoothing) the DCT spectral magnitudes which represent the expected spectral levels at specified frequencies. These samples are then geometrically interpolated (i.e. linearly interpolated in log-amplitude) to yield the expected levels at all frequencies. This simple algorithm has been referred to as "non-speech specific" since it does not take into account the dynamical properties of speech production. This adaptation is quite appropriate for speech transmission at or above 16 kbits/s, since there are enough bits to allow accurate representation of the fine structure of the DCT spectrum - in particular, the voice pitch information. At 8 kbits/s, the pitch information is no longer sufficiently preserved, and as a consequence, the received signal is degraded by a very perceptible "barbling" distortion.

A more appropriate algorithm for lower bit rates is a more complex, "speech specific", adaptation algorithm which utilises the traditional model of speech production to predict the DCT spectral levels. The prediction therefore involves two components, as illustrated in figure 3.39. The first is associated with the formant spectral envelope, and the second with the harmonic fine structure of the spectrum. Because this spectrum modelling is basic to vocoder techniques, this class of ATCs has been referred to as Vocoder-Driven ATCs. The vocoder model is used only to "steer" the ATC in its adaptation - not to code the signal. Figure 3.40 shows a snapshot in time of the input signal spectrum, the dynamic bit assignment and the receiver reconstruction for a digital coding rate of 8 kbits/s produced by the ATC.
d) **Noise Shaping**

The bit assignment prescribed by equation (3.69) produces quantization noise with flat spectral characteristics. Such characteristics are, however, known to be perceptually suboptimum. The distribution of quantization noise across frequency can be controlled by a modified bit assignment rule:

\[ b_k = \delta + \frac{1}{2} \log_2 \left[ \frac{w_k a_k^2}{D^k} \right] \quad (3.72) \]

\[ k = 0, 1, \ldots, N-1 \]

where \( w_k \) represents a positive weighting.

For a uniform weighting, the noise spectrum is flat. The noise spectrum can also be made to follow the input spectrum giving a constant SNR as a function of frequency. This noise spectrum can be varied between these two extremes which can be determined by an audibility criterion.

3.4.1.4 **Discrete Fourier Transform (DFT) Methods**

The Discrete Fourier Transform is a method typically employed in bandwidth compression systems operating in the frequency domain. This is in spite of the fact that the DFT block transform is
considered to be suboptimum in that the transform coefficients are not fully decorrelated, which is unlike the case of the KLT or DCT (ref 53). One reason for using the DFT is that its transform coefficients have a direct correspondence with real frequencies which can be of great assistance to the designer when interpreting the subjective performance of such a frequency coder (e.g. over which frequency ranges should the coding accuracy be greatest - this may or may not correspond to those regions of peak energy). It is also known that the DFT approaches the KLT performance for large block sizes \( N \). (ref 53) The DFT is signal-independent and fast Fourier transforms (FFT) algorithms do exist.

One example of a coding scheme (ref 54) based on the analysis of speech signals using the DFT is by selecting the dominant components of the transform. This method involves sampling of the speech signal at the Nyquist rate and calculating the DFT, given by

\[
X_k = \frac{1}{N} \sum_{n=0}^{N-1} x_n \exp \left\{ -\frac{2\pi ki}{N} \right\}
\]

for \( i = 0, 1, \ldots, N-1 \)
and \( k = 0, 1, \ldots, N-1 \)

(3.73)

where \( \{x_i\} \) is the input time sequence and \( \{X_k\} \) is the sequence containing the complex DFT coefficients.

The dominant spectral lines are then selected according to a local maxima algorithm and transmitted together with the spectral location information. The complex components and spectral locations for each
of the spectral lines can be transferred to the synthesizer in
digital form by log-PCM. At the receiving end, reconstruction is
performed by inserting zero valued components between the selected
components and then inverse transforming using the IDFT viz

\[ x_i = \sum_{n=0}^{N-1} x_k \exp \left\{ \frac{2\pi ni}{N} \right\} \]  

for \( i = 0, 1, \ldots, N-1 \)

and \( k = 0, 1, \ldots, N-1 \)

where \( \sim \) denotes recovered sequences.

At the transmitter, the DFT analyser operates on blocks of \( N \) samples
and its output consists of \( N \) complex components. Only \( N/2 \)
components are of interest as the input samples are real therefore
the other \( N/2 \) components are symmetrical about dc . To select the
dominant components of the transform, \( n \) lines from the spectrum
representing the highest local maxima of the spectral envelope are
selected. The real and imaginary components are quantized
logarithmically with \( r \) bits each. The optimum combination of \( N, n \)
and \( r \) was determined by subjective listening tests.

For speech bandlimited between 300 to 3000Hz and sampled at 8kHz,
the optimum block length was found to be 16m5 with a selection of
1/32 of the transform coefficients and using 4 bits of accuracy for
each component of each transmitted spectral line. This corresponded
to a compression ratio of 1:16 with a syllabic intelligibility at
75% or higher. However, the reconstructed speech did, from time to
time, contain random tones which sounded like a xylophone. This
effect is most probably due to the fact that zero valued samples in the Fourier domain are frequency slots (or holes) that move about from block to block and which are responsible for the 'musical tones'.

A. Discrete Short-Time Fourier Transform

The short-time Fourier transform in discrete form can be defined as (ref 41)

\[ X_k(j\omega) = \sum_{i=-\infty}^{\infty} h(k-i) x_i \exp(-j\omega k) \]  

(3.75)

where \( \{x_i\} \) represents samples of the input signal and \( h(k-i) \) represents a real window which reflects the portion of \( x_i \) to be analysed. The short-time Fourier transform is a function of two variables in the discrete time index 'i' and the continuous frequency variable \( \omega \). It can be interpreted in two convenient ways, either in the filter bank sense or a block Fourier transform sense. In the filter bank interpretation, figure 3.27, \( \omega \) is fixed at \( \omega = \omega_0 \) and \( X(j\omega) \) is viewed as the output of a linear time-invariant filter with an impulse response \( h(i) \) excited by the modulated signal \( x_i \exp(-j\omega_0 i) \). That is

\[ X_k(j\omega_0) = h(i) \ast \left[ x_i \exp(-j\omega_0 i) \right] \] 

(3.76)
Within this context, \( h(i) \) determines the bandwidth of the analysis around the centre frequency \( \omega_0 \) of the signal \( x_i \) and it is referred to as the 'analysis filter'.

In the block Fourier transform interpretation the time index 'i' is fixed at \( i=i_0 \) and \( x_{i_0}(j\omega) \) is viewed as the normal Fourier transform of the windowed sequence \( h(i_0-i).x_i \). That is

\[
X_k(j\omega)_{i=i_0} = \text{FT} \left\{ h(i_0-i).x_i \right\}
\]  

(3.77)

In this context, \( h(i_0-i) \) determines the time width of the analysis around the time instant \( i=i_0 \) and it is referred to as the analysis window.

The signal \( x_i \) can be recovered from its short time spectrum by means of a general synthesis equation or inverse short time Fourier transform of the form

\[
x_i = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sum_{k=-\infty}^{\infty} f(i-k) X_k(j\omega) \exp(j\omega i) \, d\omega
\]

where the sequence \( f(i) \) is referred to as the synthesis 'filter' or the synthesis 'window' as before.

It can be shown (ref 44) that the DFT (or FFT) may be used to render systems, based on the discrete short-time Fourier transform,
computationally more efficient. This notion is also extended to implementing the digital phase vocoder (ref 43) using the FFT algorithm (also mentioned in section 3.3.11).

3.4.1.5 Frequency Coding based on the Phase Spectrum

This explores the possibility of bandwidth compression by transmitting only the power of each frequency component of a signal and not the phase; a theoretical maximum of 2:1 compression should be achieved by this process. The notion of phase rejection is commensurate to the lack of phase sensitivity of the human perception mechanism. (refs 3,33)

In this process, the speech time signal is divided into contiguous blocks, where each block is weighted by a window function and Fourier transformed by the FFT algorithm. The discrete power spectrum is transmitted directly while the discrete phase spectrum is processed by a number of various methods (ref 56).

a) Co-Phase Processing.

In this case the discrete phase spectrum is discarded. At the receiver, the IDFT is taken of the power spectrum only.

\[ x_t = \sum_{k=0}^{N-1} |x_k| \exp \left( j \frac{2\pi k_i}{N} \right) \]  

\[ (3.78) \]
The initial value of $\hat{x}_i$ is:

$$\hat{x}_0 = \sum_{k=0}^{N-1} |x_k|$$

Now since the amplitude spectrum will have all positive values, $\hat{x}_0$ will be the maximum value of the function $\hat{x}_i$. Similarly for $k=N-1$ (i.e. the last sample in the block), the exponential term will be large making the value of the summation large. The shape of the resulting processed block will peak at the block edges. This peak will correspond to the join between successive segments and will thus result in a periodic peak in the reconstructed waveform. This is audible as a strong periodic click occurring at the block segmentation period. The intelligibility of the received signal is almost unaffected but the quality is seriously worsened. The principle now is to gradually re-introduce information related to the phase spectrum with minimal increase in transmission overheads by using refinements to the above process.

If each component of the amplitude spectrum takes the sign of either the real or imaginary part of that component in the complex spectrum, then the periodic peaks are reduced in the processed waveform. However, if instead only the real or imaginary components are retained for all segments then this has the undesirable consequences of 'fading' in the output signal power when the original signal has a larger quadrature component than the phase direction transmitted.
b) **Phase Quantizing**

In this case, the phase plane is divided into a discrete number of directions and allows the spectrum of each segment to assume only one of the quantized phase directions. The direction chosen is that which gives maximum power level for the sum of all components in that phase direction. The subjective improvement for a $1^\circ$ quantum angle was reported to be quite marked but harshness characterised by fixed block processing was still apparent.

c) **Pitch Synchronous Processing**

This produces less unwanted components into the power spectrum compared to methods involving segmentation into arbitrary fixed lengths, as the assumption of signal periodicity upon which block analysis is suited is better satisfied.

d) **'Continuous' Phase Processing**

Weinstein (ref 57) has considered using transmission economy in scrambling systems again by setting the whole of the phase spectrum to zero for the original voice signal. At the receiver, an artificial phase function is derived as follows.

The phase progression of a spectral component, $k$, from one sample block to the next should be roughly proportional to the
frequency-time product. For a non-dispersive channel, the phase shift is

\[ \phi_k = \omega_k NT \text{ , where } NT \text{ is the block duration} \quad (3.79) \]

\[ = 2\pi f_k \frac{N}{F_s} \]

where \( F_s = \text{sampling frequency} \)

\( N = \text{number of samples per block} \)

For the DFT sequence, \( k = \frac{f_k}{F_s} \cdot N \)

\[ \therefore \phi_k = 2\pi k \quad (3.80) \]

If \( \theta_{k,j} \) is the current phase of sample \( k \) in block \( j \), then

\[ \theta_{k,j+1} = \theta_{k,j} + \phi_k = \theta_{k,j} + 2\pi k \]

It can be seen that \( \theta_{k,j+1} = \theta_{k,j} \) i.e. the phase function is unaltered due to the cyclic nature of the phase function (periodic in \( 2\pi \) radians).

If however, block overlapping is used, (say \( p\% \) overlap, then equation (3.80) becomes

\[ \theta_{k,j+1} = \theta_{k,j} + 2\pi k \left( 1 - \frac{p}{100} \right) \quad (3.81) \]
Unfortunately, overlapping blocks cause a loss in transmission economy.

When Weinstein employed this latter method, a heavy periodic distortion in the output signal had still been reported. This is possibly due to the fact that a phase progression from one block to the next represents a running sinewave situation which cannot adequately represent a quasi-periodic signal.

In conclusion, it has not been possible to achieve the ideal 2:1 compression of the speech bandwidth due to the phase redundancy without loss in quality. It appears that some phase information must be retained in the processed speech segments. The quality is greatly improved by phase quantizing successive spectra but the associated computing complexity is high. Also pitch synchronous co-phase processing yields less unwanted components in the resultant power spectrum by segmentation.

Further research (ref 58, 59) is to be carried out to capitalise on the elusive but highly desirable goal of the ear's insensitivity to phase distortion.

3.4.2 Time Domain Coding of Speech (ref 7)

3.4.2.1 Adaptive Predictive Coding (APC)

Predictive coding systems discussed previously with DPCM have been limited to linear predictors with fixed coefficients. However, due to the non-stationary nature of speech signals, a fixed predictor
cannot track the signal values efficiently at all times. Thus the predictor must vary with time to cope with the changing spectral envelope of the speech signal as well as with the changing periodicities in voiced speech.

Adaptive prediction of speech is done most conveniently in two separate stages; a prediction exploiting the correlations between successive speech samples (or, equivalently, the non-uniform nature of the short-time spectral envelope of speech signals), and another prediction exploiting the quasi-periodic nature of voiced speech (or, the spectral fine structure representing the harmonic lines of the voiced pitch).

a) **Prediction based upon Spectral Fine Structure**

A simple method of predicting the present value of a periodic signal is to equate it to the value of the signal one or more periods earlier. For speech, the predictor has to provide some gain adjustment as well to account for amplitude variations from one period to another. The predictor can be characterised in the z-transform notation by

$$ P_d(z) = \beta_c z^{-M} $$

where $M$ represents a relatively long delay in the range 2 to 20mS and $\beta_c$ is a scaling factor.
In most cases the delay would correspond to a pitch period (or possibly, an integral number of pitch periods). The amount of prediction depends on the correlation between adjacent pitch periods. The delay $M$ is chosen so that the correlation between speech samples delayed $M$ samples apart is highest. The parameter $\beta_c$ is given by

$$\beta_c = \frac{\langle x_i \ x_{i-M} \rangle}{\langle x_i^2 \rangle}$$  \hspace{1cm} (3.83)

where $x_i$ is the $i$th speech sample and the $\langle,\rangle$ indicates the averaging over all the samples in a given time segment during which the predictor is to be optimum.

b) Prediction based upon Short Time Spectral Envelope

The short-time spectral envelope of speech is determined by the frequency response of the vocal tract and for voiced speech by the spectrum of the vocal chord sound pulses. The predictor can be characterised in the z-transform by

$$P_s(z) = \sum_{n=1}^{p} a_n z^{-n}$$  \hspace{1cm} (3.84)

where $z^{-1}$ represents a delay of one sample interval and $a_1 \ldots a_p$ are $p$ predictor coefficients. The value of $p$ is typically 10 for speech sampled at 8kHz. The predictor coefficients $a_n$ are determined such that the power of the prediction residual is
minimised in an adaptive manner over frame lengths whose duration may vary from 10 to 30mS (a time interval within which the vocal tract shape can be assumed to be nearly stationary).

c) **Combining the Two Types of Prediction**

The two types of prediction can be combined serially, in either order to produce a predictor operator \([1 - P(z)]\) that is essentially the product of \([1 - P_d(z)]\) and \([1 - P_s(z)]\). The prediction with \(P(z)\) can realise a higher order prediction gain than it is possible with either \(P_d\) or \(P_s\) acting alone. However, the total prediction gain \(^1\) (in dB) is not the sum of the gains for \(P_d\) and \(P_s\) acting singly on the speech signal. Used in combination, the first predictor achieves the bulk of the gain (typically 13-14dB), and the second predictor (now operating on a signal that is less predictable than the original speech) achieves the balance (typically another 3dB).

D. **Perceptual Criteria for Optimising Predictive Coders** (ref 7)

To ensure that the distortion in the speech signal is perceptually small, it is necessary to consider the spectrum of the quantization noise and its relation to the speech spectrum. The theory of auditory masking suggests that noise in the frequency regions where

\(^1\) The prediction gain is the SNR Gain of a DPCM system employing the predictor over a similar bit-rate PCM system.
speech energy is concentrated (such as formant regions) would be partially or totally masked by the speech signal. Thus, a large part of the perceived noise in a coder comes from the frequency regions where the signal energy is low (see section 3.4.1.3 on ATC noise shaping). Furthermore, what needs to be minimised is not the power of the quantization noise but its subjective loudness. Figure 3.41 shows a predictive coder capable of producing any desired spectrum of the quantizing noise. The filter $F$ redistributes the noise power from one frequency to another frequency. This filter can be used to reduce the noise in the regions where the signal is low while increasing the noise in the formant regions where the noise could be effectively masked by the signal. If $F=P$; the result is minimum unweighted noise power. The spectrum of the output noise is white producing a very high SNR at the formants, but a poor SNR in between the formants. If $F=0$ no feedback implies that the output noise has the same spectrum as the original speech, but is shifted downward - a good choice if our ears were equally sensitive to quantizing distortion for all frequencies. An intermediate design is exemplified by

$$F = P(a_p z^{-1})$$

(3.85)

where $a_p$ is an additional parameter introduced to increase the bandwidths of the zeros of $1-F$. The increased bandwidth causes the noise to peak in the format regions accompanied by a reduction in noise in regions where the signal level is low. An example of the envelope of the output noise spectrum together with the corresponding speech spectrum is shown in figure 3.42.
The generalised predictor shown in figure 3.41 can sometimes be preceded with a signal pre-emphasiser to boost the high frequencies prior to analysis. The emphasis filter has a transfer function of $[1-0.4z^{-1}]$. If an adaptive 3-bit quantizer is selected to be optimum for a Gaussian PDF and with a proper choice of $F$, the output speech quality is subjectively close to that of 7-bit log-PCM, which has an SNR of 33dB.

3.4.2.2 **Time Domain Harmonic Scaling (TDHS)** (refs 2,60,61)

Frequency scaling of speech signals by methods based on short-time Fourier analysis (STFA), analytic rooting and harmonic compression using a bank of filters, is a complex operation which requires a large amount of computation. By incorporating pitch frequency information into a frequency scaling process based on the short-time Fourier analysis, it is possible for a good approximation to perform this scaling in the time domain with only a few arithmetic operations.

The time-domain harmonic scaling algorithm consists of weighting several adjacent input signal segments (with pitch dependent duration) by a suitable window function, to produce an output segment. In the frequency domain, the time-domain operations are equivalent to the spectral shifting of the individual pitch harmonics depending upon the centre frequency of the sub-band in which each harmonic component is located. The number of sub-bands into which the speech band is divided is pitch dependent. Time scaling is achieved by properly choosing input and output sampling rates.
The TDHS algorithm is based upon the assumption that the fundamental frequency $F_0$ (the pitch) of the input voiced-speech signal is approximately known.

The algorithm is developed using STFT and a contiguous filter bank analysis of a signal $x(t)$ (also used in section 3.3.11 regarding the phase vocoder). The window function constitutes the impulse response of the lowpass prototype filter of the filter-bank having centre frequencies at $\omega_k$, $k = 1, 2, ..., L$ and impulse responses according to:

$$h_k(t) = 2h(t) \cos \omega_k t$$  

(3.86)

The bandwidth of each filter is $\Delta \omega$ i.e. each filter spans the range $\omega_k - \Delta \omega/2$ to $\omega_k + \Delta \omega/2$. The frequency scaling of the input signal $x(t)$ by a factor $q$ can be achieved by multiplying the carrier frequencies as well as the phase derivative of the Short Time Fourier Transform (STFT) by that factor. The $\omega_k$'s are chosen such that $\omega_k$ is equal to $2\pi F_0$. If there is an error in the estimation of $F_0$, then the error should be small enough such that each spectral line of the signal harmonic will be located in a separate sub-band, $\Delta \omega$.

Considering the traditional spectral model of a quasi-periodic voiced sound sketched in figure 3.43(a), the spectrum reflects the pitch periodicity of the sound by finite width 'teeth' spaced at the fundamental pitch frequency. The amplitude of each tooth
conditioned by the resonance structure of the vocal tract and the simultaneous modulations of $|X(\omega_k,t)|$ and $\phi(\omega_k,t)$ from equation 3.55 influence the non-zero width between the spectral teeth. Studies with channel vocoders indicate (ref 46) that the band occupancy of the amplitude modulation is of the order of ±25Hz and similar studies with the phase vocoder indicate comparable occupancy for frequency modulation. If the $\omega_k$'s are linearly compressed by $1/q$ then the teeth may overlap as shown in figure 3.43(b). Frequency division incorporating scaling of the phase derivative also causes narrowing of the modulated teeth to prevent spectral overlap [figure 3.43(c)]. The frequency division factor is still limited to 3 or less such that spectral overlap of the harmonic teeth should be prevented.

The time domain harmonic scaling algorithm is derived by first considering frequency scaling via the short time Fourier transform (i.e. frequency scaling the filter bank signals as in the case of the phase vocoder (ref 40)) then by using the inverse transform, the operation is developed in the time domain (ref 2). The frequency division process is based upon combining samples of the time waveform that are pitch intervals apart and weighted by a suitable window function (a triangular window was found to be practical in this case).
The frequency division process is ordered according to the expression

\[ y^{1/c}(\ell) = x(\ell) \cdot h_N(\ell) + x(\ell-N_p) \cdot h_N(\ell+N_c) \]  

for \( \ell = 0,1,\ldots,N_c-1 \)  

(3.87)

where \( x(\ell) \) is the input time sequence,

\( y^{1/c}(\ell) \) is the frequency compressed output time sequence

\( h_N(\ell) \) is the window weighting function of length \( N_c \) samples

\( N_p \) is the current pitch interval of \( x(\ell) \) (No. of Samples)

and \( c \) is the compression factor.

Likewise, frequency multiplication is ordered according to the expression

\[ y^S(\ell) = x(\ell-N_p) \cdot h_N(N_s-\ell) + x(\ell-2N_p) \cdot h_N(2N_s-\ell) \]  

(3.88)

where \( x(\ell) \) is the input time sequence

\( y^S(\ell) \) is the frequency expanded output time sequence

\( N_s \) is the number of samples contained in the window

\( N_p \) is the number of samples in the pitch interval of \( x(\ell) \)

and \( s \) is the frequency expansion factor.

The applications of this process include bandwidth compression and expansion which according to the above expressions require accurate pitch tracking. This is because in the derivation of these operations, it was assumed that the pitch harmonics are specified
accurately in frequency by the analysis filters. This tolerance can be relaxed if provision is made not only to allow for the amplitude variations of the pitch harmonics but also the phase derivative of the deviation of these harmonics from the centre frequency of the analysis filters as does the phase vocoder.

One disadvantage of the frequency scaling process in the time domain is that the compression/expansion is linear across the whole of the speech band. Perhaps additional considerations may be included to use piece-wise frequency scaling according to perceptually significant frequency regions of the speech band as does the sub band coder and the adaptive transform coder.

The TDHS process has been demonstrated (ref 60) as a pre-processor in conjunction with digital coders, (figure 3.44), with quite favourable results i.e. the performance of a CVSD coder working at 7.2 kbits\(^{-1}\) can be made to approximate that of a coder operating at twice the bit rate without a THDS pre-processor.

Other digitisers such as sub band coding and adaptive transform coding have also been preceded by the TDHS processor. (ref 61) The sub band coder combined with TDHS at 9.6 kbits/s was reported to provide a quality equivalent to that of SBC alone at 16 kbits/s, thus a bit-rate advantage of 6.4 kbits/s was realised. When the combined adaptive transform coder and the TDHS system was subjectively evaluated, a bit-rate advantage of 4 kbits/s at 7.2 kbits/s was achieved. The SBC plus TDHS method was noted as the most attractive since its quality was comparable with ATC plus TDHS at 16 kbits/s yet it had a lower complexity than the ATC plus TDHS system.
Alternative applications of this process include time scale variation of speech by replaying the frequency scaled signal at a \( \frac{1}{c} \) (or \( \frac{1}{s} \)) speed factor so as to restore the frequency distribution back to its original position after TDHS processing. This can be implemented as an on-line process for "speeded-speech" for the blind, slowed down speech for language teaching or time scale normalisation of discrete speech utterances for isolated word recognition systems (IWRS).

3.4.2.3 Time Encoded Speech (ref 62)

This method operates principally by encoding the time interval between speech events (hence the reference to 'time-encoded speech'). It depends upon the transmission of coded shape descriptors for successive segments of the speech waveform. This can be considered as the time domain counterpart to frequency pattern matching vocoders (section 3.3.1.1(c)). The waveform can be broken into these segments between successive real zero crossings\(^1\) of the function. For each such segment of the waveform, the code consists of a single digital word. This word is derived from two parameters of the segment; its quantized time duration and its shape. The shape may be compared to a 'catalogue' of shapes and a code selected which identifies the shape in the catalogue nearest to the actual segment-shape. One strategy that is considered is to

\[^1\] A real zero crossing occurs when the excursion of \( x(t) \) actually crosses the time axis, which if occurs at time \( t_1 \) then \( \text{sgn}(x(t_1^-)) = -\text{sgn}(x(t_1^+)) \).
classify wave segments on the basis of the number of complex zeros\textsuperscript{2}, (see figure 3.45), and these are converted to real zeros by single differentiation. The large number of naturally occurring shapes (symbols) can be mapped into a smaller number. With a 23-symbol alphabet the reconstructed speech gave good intelligibility and speaker identification.

Reconstruction of time encoded speech is performed by reproducing stored segment forms in the correct sequence at the correct duration. Since the mean amplitude is a relatively slowly varying quantity, after every eight transmitted symbols, an additional symbol is also transmitted to indicate the mean speech amplitude over the eight proceeding symbols. This also ensures interword muting.

The transmission bit-rates tested for this system are 3600-5400 bits/s.

The effects of channel errors have also been tested on this system in which significant degradation of speech quality is perceived when 1 symbol in 8 is corrupted (corresponding to a ber of 3 in $10^{-2}$, i.e. a comparatively severe test).

\textsuperscript{2} A complex zero crossing is said to occur when the slope of $x(t)$ changes direction without crossing the time axis, which if occurs at time $t_2$ then $\text{sgn}(\dot{x}(t_2^-)) = -\text{sgn}(\dot{x}(t_2^+))$. 
3.4.3 Comparison of Speech Coders

3.4.3.1 Transmission-Rate Profile (ref 7)

Figure 3.46 shows a "digital-spectrum" of speech coding rates of interest. The figure highlights the dichotomy between non-speech-specific waveform coders (discussed in section (3.2) that need relatively higher transmission rates, and speech-specific vocoders (section 3.3) for digitisation at relatively lower bit rates. The figure also indicates the quality of speech reproduction that can presently be attained at a prescribed bit rate. The quality characteristics are denoted as commentary, toll, communications and synthetic.

The term 'toll-quality' is used to imply quality comparable to that of an analogue speech signal having approximately the following properties:
Frequency range = 200 to 3200Hz; Signal-to-noise ratio ≥ 30dB; Total harmonic distortion ≤ 2-3%.

As figure 3.46 suggests, one can achieve telephone toll quality for speech signals at coding rates of 16 kb/S. At bit rates exceeding 64 kb/S, it is possible to obtain the SNR's and harmonic distortion characteristics of toll quality with input signal bandwidths significantly wider than normal telephone speech (e.g. 0 to 7kHz or better). This grade of quality is loosely referred to as commentary quality. It is useful adjunct for digitising some varieties of broadcast material.
At rates below 16 kb/s and specifically the data speed range 9.6 to 7.2 kbits/s the waveform coders discussed in section (3.2) can provide communications quality speech. It is noted that the signal is highly intelligible, but has noticeable quality reduction.

Coders in the source coding (analysis-synthesis) range of 4.8 kbits/s and below provides synthetic quality, where the signal loses substantial naturalness and sounds automation-like. Talker recognition is substantially degraded and the coder performance is talker dependent.

3.4.3.2 Characteristics of the Coders

The characteristics of the types of speech coder described hitherto are summarised in Table (3.4) (ref 17). The entries for speech quality and complexity are intended only as a relative guide.

It can be seen that there is a large variety of speech coding methods available for the system designer to make the best compromise between the 3 conflicting objectives of good speech quality, low transmission rate and low equipment complexity. A more difficult problem is to assess also how faults in the reproduction of the transmitted speech can affect the communicators; this is not, however, considered in this thesis.
3.5  Wideband Speech Processing

Having described systems for conveying toll quality speech over analogue or digital channels, we now extend our review to the application of these systems by invoking further development where necessary with the principal task of compressing a wide band speech signal (0-7kHz or greater) into a telephone bandwidth (or equivalent bit rate) channels. In the second part of this section, the possibility of improving the quality of speech received from a commercial telephone channel is also briefly reviewed.

3.5.1 Voice-Excited Vocoder (ref 47)

Voice excited vocoders (or baseband coders) were originally proposed to circumvent the pitch tracking problem (see section 3.4.1.1) but here, they are recalled to apply them to wideband speech compression.

The types of sound that are degraded the most when conveyed via a telephone channel (0.3-3.4kHz) are not the voiced sounds as most of the spectral energy resides within the telephone band. Unvoiced fricatives and affricatives have a predominant amount of spectral energy above the normal telephone range and it is these such sounds that the VEV and other wideband systems would sensibly endeavour to improve over those sounds conveyed over the telephone channel.

It is also noted that the ear is not very sensitive to spectral modifications of these sounds. For example, it has been reported in earlier experiments (ref 47) that a broad resonance around 3kHz
excited by noise sounds respectively like a /ʃ/ and the subjective perception does not depend on the width of this resonance within wide limits. The same is true for the /s/ which can be synthesized from a band of noise around 5 to 7kHz.

Using the above premise, it is considered that the band between 3 and 10kHz can be compressed into a few hundred Hz using the same channel vocoder method shown in figure 3.30.

3.5.1.1 The Spectral Flattener

Speech synthesis by the vocoder requires a flat spectrum of constant power density and of the proper type (discrete or continuous). The transformation from the shaped and band limited spectrum of the uncoded speech band to a flat wideband spectrum of the proper type between 3 and 10kHz can be achieved by non-linear distortion. The non-linear distortion method 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 3.48, which resembles the character 'W'. The multiplier transfer function produces many zero crossings and distortion components up to 10kHz and beyond. This 'W' function spectral flattener is used for signals containing both voiced and unvoiced excitation without explicit pitch detection or voiced-unvoiced decision. The W-function does however spread 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 combination of the logarithmic compressor and the 'W' transfer function does indeed produce flat distortion components between 3 and 10kHz, as required, for all types of input sounds.

The complete vocoder system is shown in figure 3.49. A 3.2kHz uncoded baseband is incorporated along with nine vocoder spectrum channels covering the band from 3.2 to 10kHz. The total transmission bandwidth is less than 3.5kHz. A possible allocation of bandwidth for the nine channel version is shown in figure 3.50; the speech band from 80 to 2000Hz is shifted upward by 700Hz to accommodate the speech components between 80-300Hz. The shift makes space for the nine vocoder channel signals which carry the intelligence in the band 2-10kHz. The channels representing the speech frequency components immediately below 10kHz are located near the lower edge of the transmission band. Thus, if the transmission channel has a lower cut-off frequency above 300Hz, only the very highest frequencies in the vocoder multiplex transmission would suffer and no 'holes' in the spectrum would be created. Low index FM or SSBSC are possible multiplexing methods. The effective compression for the entire transmitted band is 9920Hz into 2370Hz i.e. 4.2 : 1.

Informal listening tests indicate that this scheme yields a net gain in both intelligibility and quality over a 3.5kHz band of speech. Critical listening tests using isolated English words indicate a considerable gain in articulation compared to a simple 3.5kHz
circuit; the improvement is particularly noticeable in words containing plosive and fricative sounds. Listeners commented that the vocoder output is much more similar to the wide band original speech than the 3.2kHz band limited sample.

The notion of applying voice excited vocoder techniques to compress wideband signals into the telephone bandwidth is further considered (ref 63) where the base band signal occupies 2.5-3.5kHz and the remaining band 2.5-7kHz is divided into 6 or 12 bands, rectified and time division multiplexed together by a sampler switch as shown in figure 3.51. The TDM signal containing the vocoded channels is SSB modulated and added to the base band signal for transmission. At the receiver, the base band and vocoder band are separated by filters, the vocoder band is SSB demodulated and fed to a switch operating in synchronism with the transmitter switch. The vocoded channel signals are used to control fixed frequency oscillators each set to the centre frequency of its respective channel. The modulated signals are added to the base band signal to form the reconstructed output signal. It is stated that an improvement is observed over a normal 3.5kHz signal by means of the PICOR (pilot controlled overtime reproduction) method with a 3.5kHz base band and artificial 'overtones' up to 7kHz.

This latter method perhaps may be not so successful when attempting to reproduce fricative and affricative sounds as with Schroeder's VEV. In the case of PICOR, the high frequency band is produced by a series of oscillators, whereas in the case of Schroeder's VEV, the
high frequencies are produced by extending the fine structure of the base band (using a non-linear transfer function) which are of course related to the spectral structure in that base band.

3.5.2 Digital Compression of Wideband Speech Signals

The BBC (refs 64-67) have carried out research in the area of digital sound compression; initially they considered a type of companding used in PCM signals which is applied to high quality sound (music and speech). With linear quantizing (i.e. no companding) of 13 bits/sample at a sampling rate of 32kHz, the sound signal is transmitted at a basic rate of 416 kbits/s per programme channel. "Near instantaneous" companded quantization (n.i.) can be used to offer a bit rate reduction. The nicamp quantizer codes a batch of say 24 samples choosing a linear quantizer that will just accommodate the largest sample value in the batch. If 4 such linear quantization laws are available, then a 2-bit scale-factor word is transmitted once per batch. Using this type of compander, it has been shown that it is possible to reduce the number of bits per sample to about 10 using the 'near instantaneous' companding action, the programme quality (comprising of speech and music) being virtually indistinguishable by critical listeners from that of a linearly-coded system, employing 13 bits per sample with dither (section 3.2.1.3). The A-law of instantaneous companding which also requires 10 bits per sample results in inferior quality and is not considered to meet BBC requirements for high quality sound transmissions.
The BBC (ref 65) have further considered sending programme speech contributions digitally over the public telephone network by devising equipment which could be used to replace the telephone handset and to generate a signal which would be less susceptible to the impairments of the communications channel.

To transmit a digital signal via the telephone network at a maximum rate without intersymbol interference, some form of correction for both amplitude and phase distortion must be used. A device that modulates the digital signals considered in combination with a complementary device that demodulates them at the receiving terminal - possibly also providing some degree of correction for the impairments of the circuit is called a 'modem'. Commercial modems are available for giving bit-rates up to 9.6 kb/s over suitable telephone circuits with adaptive equalisation.

In the telephone network, it is not yet convenient for the customer to access 64 kb/s digital channels. The Datel 48K service (where 12 channels of an entire group is leased) giving 48 kb/s is expensive due to the number of telephone channels displaced. It appears that if the speech signal is to be digitized prior to transmission via the telephone channel then speech coding devices operating at 1.2 kb/s using modems with no equalisation or 9.6 kb/s using modems with adaptive equalisation need to be considered. If either waveform coders or vocoders at these bit rates is used then the quality and intelligibility of the received speech was found to be quite unacceptable for the application required (i.e. broadcast quality
speech contributions). A number of speech coding systems (discussed in section 3.2 and 3.3) had been considered, also the possibility of exploiting the silence pauses in speech but this technique was shown to be not practicable as the storage required would be too great and the benefits were small. It was concluded (ref 65) that no system of coding and transmission is readily available for providing speech circuits at the bit rates required that would give a speech quality better than that normally achieved using an ordinary telephone handset.

With the possible availability of customer access to 64 kbit/s capacity channels, research has continued in order to transmit commentary grade (7kHz) and with 64 kbit/s coding. (refs 65-68) The methods that are considered are as follows.

3.5.2.1 Separate Description of the Envelope and Zero-Crossing of the Top Octave Band (ref 66)

In this method, the envelope and zero crossings of the top-octave band of the signal can be isolated and separately coded for transmission. The rate of change of envelope in the upper band (3.4-6.8kHz) is such that the envelope can be transmitted in a bandwidth of approximately 750Hz. However, no clear method of encoding the top-octave zero-crossings had been conceived as the zero-crossing signal cannot be sufficiently band limited to afford accommodation of the coded envelope, zero-crossing signal and the coded version of the lower sub-band within the 64 kbits/s
available. Perhaps instead of coding the number of zero-crossings, the rate of change of this number may be a better candidate for transmission.

3.5.2.2 Sub-Nyquist Sampling (ref 66)

To avoid unwanted alias components of a sampled signal overlapping the wanted base band signal, the sampling rate should be at least twice that of the highest frequency in the base band signal. However, if the required band of frequencies (or an upper part of this band) is initially comb-filtered, then it is possible to choose a particular sampling frequency such that the alias components fall in the gaps of the spectrum produced by the comb filter. In a decoder, the alias components may be removed by a second comb-filter similar to the first. The problem of using this technique is that comb-filters have a considerable deleterious effect on audio quality. It was found however, that if the action of the filter was confined to the upper two octaves of the speech signal, there is then much less effect on the sound quality. Using informal listening tests seeking a minimum effect on audio quality, it was found that the spacing between the 'teeth' of the comb filter response was best at about 1kHz.

3.5.2.3 Pitch Halving (ref 66)

It can be seen by simple inspection of the waveforms that adjacent speech waveform segments, of about 10 to 30ms duration corresponding to the pitch interval, are very often similar to each other; transmission of alternate segments only (effectively halving the
pitch) is then sufficient, provided that the transmitted segments (after restoration of the pitch) are each repeated in the receiver. In practice, of course, significant difference between blocks of samples do occur, and distortion of the output signal then arises when blocks are omitted or repeated.

This coder was then used to remove alternate 16mS segments of the signal (i.e. non-pitch synchronous blocks), the decoder then repeated each received segment. The problem with segment repetition was that the audio output becomes modulated at the repetition rate (1/16mS) causing a low frequency buzz. Higher frequency clicks occurred when there were large discontinuities between adjacent samples. However when the action of this system was confined to the signal above 1.7kHz, the signal distortion was considerably reduced. Furthermore, by using pre- and de-emphasis of the higher frequencies, before and after the coder and decoder, respectively, the audibility of the high frequency distortion can be reduced. Also a notch filter can be incorporated in the decoder which suppresses the frequencies causing gross distortion (i.e. at the block repetition rate).

For subjective testing, a recording was made of the lower frequency band using 1.7kHz low pass filter encoded at 6 bits/sample, n.i. companding to a bit rate of 24 kbits/s; and 5 bits per sample and n.i. companding, corresponding a bit rate of 40 kbits/s for the signal above 1.7kHz giving a total bit rate of 64 kbits/s. The results of the informal listening tests were encouraging, although distortion was audible with some speakers. Pitch synchronous halving was reported to give better subjective results.
3.5.2.4. Variable Sampling Rate using Constant Bit Rate (ref 66)

If an utterance is spoken which contains no components above 2kHz in frequency, then sampling at a little over 4kHz would be sufficient to fully describe that signal. If the next utterance has components up to 4kHz, then sampling at a little over 8kHz would be sufficient. If the sampling rate were to be varied in accordance with the speech-spectrum variations and the quantizing accuracy is reduced as the sampling is increased (and vice versa), then the bit rate can be kept constant.

A simulation (ref 66) of this method was made in which the variable sampling rate could have one of four values. These were 16kHz, 10.5kHz, 8kHz and 6.2kHz. The corresponding bits per sample would be 4, 6, 8 and 10, to give an approximately constant bit rate of 64 kbits/s. To determine the appropriate choice of sampling frequency, four low pass filters and three signal-level threshold detectors were used.

The results of this system showed a marked improvement in the clarity of the received speech compared with normal telephone quality, however, the main defect with this method was that it was possible to hear clearly the effect of the varying bandwidth as the sampling rate was varied. This produced variation in the spectrum and audibility of the background noise, and some observers rated these effects disturbing. Nevertheless, this method is still considered to give speech quality which may be acceptable for commentary circuits.
3.5.2.5 Sub Band Coding of Wideband Speech (ref 67)

This method again relies on the availability of 64 kbits/sec circuits accessible to the customer's premises.

The notion of sub-band coding (discussed in section 3.4.1.2) has the advantage of containing quantization noise within the band that gives rise to that noise thus preventing masking of one frequency range by quantizing noise in another frequency range. In one practical split-band coding system (ref 67), the spectrum was divided into three bands - up to 1.75kHz, 1.75 to 3.5kHz and 3.5 to 7kHz. As before, integer band sampling can be used to avoid the necessity of frequency translation by modulators.

These bands were encoded using near-instantaneous companding; the number of transmitted bits/sample was 6 for the two lower bands and 3 for the upper band. The scale factor period for the companding was 10mS for each band. The sampling rate for the two lower bands was 3.5kHz thus the transmission rate for these two bands combined was 42 kbits/s. The band from 3.5 to 7kHz was allocated 3 bits/word. With a sampling rate of 7kHz this band required a data rate of 21 kbits/s, making a total data rate of 63 kbits/s for the three bands leaving 1 kbits/s for ranging information for the three companders. The performance of the system had been assessed subjectively and the results were considered "very promising" and improved compared to conventional 64 kbits/s telephonic speech.
A two-band coding scheme was considered by Johnston and Goodman (ref 68) in which the lower band 0 to 3650 Hz is coded with 4-bit ADPCM and the upper band 3600 to 6800 Hz is coded with 3-bit or 4-bit APCM. The aim of this coding strategy is to provide a 7kHz bandwidth commentary grade speech or music transmission at 56 or 64 kbits/s. The coding scheme is illustrated in figure 3.52; here, the two bands shown could be further subdivided to produce enhanced quality at the cost of additional equipment complexity. With the 3-bit upper band coder, 3 kbits/s of framing and synchronising information can be added to provide a 56 kbits/s signal. With 4 bits/sample in the upper band, 4 kbits/s of side information can be added to produce 64 kbits/s.

Both the ADPCM and APCM coders contain robust adaptive quantizers, the step size, $\Delta$, of the quantizer with $2^B$ output levels varies according to

$$\Delta_{i+1} = M(|I_i|) \Delta_i^B$$  \hspace{1cm} (3.88)

where $i$ is the time index, $I_i$ is the $i$-th transmitted code word, $\beta$ is the leakage constant, and $M(1), M(2), ..., M(2^B)$ are the step size multipliers used in reference (68).

The performance of this system by informal listening experiences of coded speech and music suggested that the system produces quality comparable to good AM reception. With 3 bits per sample in the upper band channel, some high frequency quantizing noise was perceptible. With 4 bits per sample, the system noise was noticeably quieter.
The lower channel produces telephone quality speech at 32 kbits/s. It is therefore possible to incorporate the encoder in a service that allows the user to choose between sending one commentary-grade signal or two telephone-quality signals over a 64 kbits/s line, depending upon the nature of communication.

To improve the band separation and response characteristics imposed by the practical filters, quadrature mirror filters can be used (section 3.4.1.2).

3.5.3 Quality Improvement of Telephone Speech

This section considers enhancing the quality of telephone speech by noise reduction and bandwidth extension. The methods investigated allow no processing at the transmitter terminal, the only knowledge (or information) that the receiver has is that a telephone channel was used prior to speech reception.

Several devices (ref 69) have been investigated which may suggest a means of improving the acoustical quality of the telephone contributed speech to broadcast programmes.

3.5.3.1 Amplitude/Frequency Equalisation

The cumulative effects of frequency responses of different parts of the telephone system impair the speech quality and may often effectively reduce the pass band to considerably less than the normal 0.3-3.4kHz. It is therefore desirable to equalise the characteristic of telephone signals within the normal passband.
Equalisation is normally carried out by test signals. The only signal here is the speech signal by the caller. The spectrum of telephone-speech signals deviates by up to 20dB from the average speech spectrum whereas studio quality deviates up to only 3dB.

The practical method used for equalisation was to divide the spectrum into 10 1/3-octave bands to measure the signal spectrum. The required equalisation was found to be achieved usually by varying the depth of two fixed frequency notch filters (one at 920Hz and one at 1.4kHz). It was thus found that telephone signals could be matched to within ±3dB of the average speech spectrum. The time constant of the adaptive equalisation did prove to be critical as far as subjective effects are concerned.

3.5.3.2 Extension of Bandwidth

It is possible to retrieve some speech components from outside the telephone bandwidth by equalisation but there would be an overall signal impairment due to the consequent increase in noise. It is considered that if telephone speech signals are to be processed to become similar to high quality speech, the processing must include some means of synthesising components outside the telephone bandwidth.

a) Low Frequency Synthesis

Voiced sounds such as vowels and part of some consonants account the low frequency energy in speech signals. It is considered possible
to estimate the nature of the original speech components below 300Hz from the telephone signal where the fundamental frequency and its harmonics are synthesised by a non-linear process of full-wave rectification and bandpass filtering to 80-300Hz. The output signal contains components at the fundamental frequency plus harmonics.

The success of the rectifying technique in synthesising a realistic low-frequency spectrum depended critically on the quality of the prevailing input signal. The results improved with equalisation applied before the input to the synthesiser.

b) Higher Frequency Synthesis

The main components of speech energy at frequencies above the telephone bandwidth are unvoiced and occur during sibilants and plosives. The main speech energy components within the telephone bandwidth are voiced, therefore, because of the different proportions and nature of voiced and unvoiced components, it is necessary to differentiate between them in order to generate realistic higher-frequency speech components. Using an unvoiced/voiced detector on telephone quality speech posed extreme difficulty. The most successful method compares speech energy in upper part of the telephone band to that in the lower part but this proved unreliable (for further discussion of U/V detectors see section 3.6.1).

No simple means of generating high frequencies was discovered at this time and used to advantage and so it was considered that the best that could be done was to compensate for the lack of high
frequencies by providing a boost to components in the upper part of the telephone bandwidth. The amplitude/frequency characteristic of the high frequency booster is sketched in figure 3.53. It was reported that for most telephone calls, a 10dB boost could be applied with advantage.

The results of the subjective tests on the processing discussed so far showed that some observers indicated that they were used to listening to broadcast telephone calls and that they sometimes found that the unprocessed signals less annoying because they were able to adapt their senses to listening to them very quickly. They also compared the processed signals to very poor quality studio signal and some observers found that the slight increase in noise level, which occurred when the signal was equalised, increased their annoyance.

For some of the processed items, the synthetic low frequency signals were not considered to be very natural. The equalisation and h.f. boost increased the level of distortion products present in the original telephone signal. Overall, the subjective tests showed that a system based on the most successful combination of the devices effected only a slight improvement in the acoustical quality of the telephone calls used in the tests. Many telephone contributions which have been broadcast were apparently of poorer acoustical quality than those selected for the tests. These poorer quality calls had high noise levels and a high percentage of non-linear distortion; neither of these two impairments were treated successfully.
Further work has been carried out towards the method of high frequency regeneration (ref 70). This, however, was considered to replace Schroder's non-linear processing of the excitation signal in base band vocoders. The process considered may in fact be used with advantage in extending the bandwidth of telephonic speech. What was proposed is the duplication of the base band spectrum.

Spectral duplication can take the form of (a) spectral folding and (b) spectral translation. Assuming that the bandwidth of the resultant signal is to be $F_W$ Hz which is an integer multiple of the base band width $F_B$ such that $F_W/F_B = L$. This integer band assumption greatly simplifies the two implementations.

Figure 3.54 shows the results for $L = 3$. Figure 3.54(a) shows the base band spectrum; figure 3.54(b) shows the result of spectral folding, and figure 3.54(c) the result of spectral translation. Figure 3.55 shows the spectral folding process using integer band folding. To perform an $L$ band spectral fold, one simply inserts $L-1$ zeros between samples of the transmitted base band. This process is merely that of upsampling, which is used to produce the spectral fold. Figure 3.56 shows the general system for integer band spectral translation. The multiplication by $(-1)^t$ gives a signal with a mirror image spectrum (as with the QMF method, section 3.4.1.2). Again, upsampling is also used to produce spectral folding. The filter $H(z)$ is a multiple bandpass filter, as shown in figure 3.57 for $L=3$, which passes those bands that have the same shape as the base band, i.e. every other band. The filter $1-H(z)$ then passes the intervening bands. The sum of the outputs of the two filters constitutes the required extended spectrum.
The results of these experiments using high frequency regeneration (HFR) by spectral folding yield distortions in the form of added tones at even multiples of the folding frequency of $2F_B$ Hz. This was eliminated by subtracting the short term dc in the base band signal as the dc is normally folded into multiples of $2F_B$ Hz.

The results of the spectral translation had not yet been reported but the sound quality would be expected to be similar to those from the spectral folding process.

Another reason for the spectral tones is that with spectral duplication, the harmonic structure is interrupted at multiples of $F_B$ Hz. It therefore should be possible to eliminate the tones if the width $F_B$ was adjusted to be a multiple of the pitch fundamental frequency on a short term basis.

A further system for the extension of the bandwidth of a bandlimited signal was demonstrated (ref 71) (not yet published) by the BBC research department. The method shown in figure 3.5B, simply uses a noise generator for high frequency regeneration which is modulated by components between 2.4 and 3.4 kHz. The resultant signal is high pass filtered from 3.4kHz then added to the telephone bandlimited signal.

For low frequency synthesis, a square wave oscillator is driven in repetition frequency by a pitch detector (operating artificially from the full band input signal). The amplitude of this oscillator is multiplied by the bandlimited signal. The 0-300Hz range of components from this oscillator signal are then added to the bandlimited signal.
The informal subjective evaluation of this system yielded remarkably good results, considering the simplicity of the process involved. There was no unvoiced/voiced decision although the pitch-tracking must be adjusted to operate on the telephone bandlimited signal in the future.

3.5.3.3 Noise Reduction System

The unwanted noise generated within a telephone system contains both impulsive and random components. For telephone circuits which have a high attenuation, the final SNR is quite unacceptable for broadcasting. Equalisers also seem to increase the noise level.

The scheme as shown in figure 3.59(a) illustrates a possible noise reduction system (refs 13,69). The gain $A_1$ is derived from the syllabic components of the envelope of the input signal. If the gain $A_2$ is large, the device operates as a noise gate. With low gains of $A_2$, the device operates as a syllabic expander. Typical transfer functions of the system are shown in figure 3.59(b). It was reported (ref 69) that unless the incoming SNR was better than about 40dB, this device produced adverse effects. With some telephone calls, modulation of the noise by the signal was distracting. If the threshold of the expander was set too high, the intelligibility of the speech was impaired. It was considered that a signal analyser capable of distinguishing between speech and non-speech signals and cancelling noise by means of a correlation technique was necessary.
More success has been reported by using companders on telephone lines for audio teleconferencing (ref 13). This involves expanding the speech at the transmitter and compressing at the receiver. The companders are non-instantaneous such that the signal is not companded when speech is present, the signal is only companded during silence pauses in order to suppress background noise and impulse switching noise. If the compression is large, the telephone connections may sound 'dead' and it is difficult to perceive that a telephone connection exists. Another effect is an increase in background noise apparent at the beginning of every spoken sentence. This is more of an annoyance rather than an effect on intelligibility.

A noise cancellation circuit has also been used (ref 13) to subtract unwanted background noise from the speech signal. The circuit comprises a trough detector to sense the background noise level and a differential amplifier. The former is essentially a peak detector with the diodes reversed and thus has decay time allowing it to track a fall in a peak detector output. The rise time however is made very large so that only long term changes are followed; a value of 20 seconds was found suitable. This form of cancellation is also efficient for periodic noise sources (ref 13).

At present, as far as enhancement of wideband processing is concerned, no attention has been given to separating higher frequency from low frequency speech utterances namely voiced and unvoiced signals. It may thus be more advantageous to concentrate enhancement for unvoiced sounds only leaving voiced sounds unprocessed – especially with moderate SNR telephone signals.
This notion directs us to briefly investigate the field of voiced/unvoiced decision which may be used prior to telephone speech enhancement which is the topic of the next section.

3.6 Voiced/Unvoiced Detection

The Voiced/Unvoiced/Silence (V/UV/S) decision forms a critical role in some of the systems already discussed and will be also used for some of the notions to be later explained in this study. This decision is significant insofar as it forms the front-end of systems upon which later decisions will be based. It will be apparent that the accuracy of the V/UV/S detection can be quite high when the input signal is wideband, however, for telephone speech inputs, the accuracy of the classification degrades quite significantly (ref 72).

3.6.1 Telephone Quality Speech

In the paper (ref 72) by Rabiner et al, the evaluation of a statistical approach to Voiced-Unvoiced silence analysis for telephone-quality speech is considered. A large number of parameters (70) were included in the investigation, including 12 LPC coefficients, 12 correlation coefficients, 12 PARCOR coefficients. Many of the parameters were immediately eliminated because they provided almost no separability between the three decision classes. The remaining parameters were used in a optimisation process to determine the five best parameters to use for a voiced-unvoiced-silence analysis. It was found that the particular parameters set
effective for wideband inputs was not equally effective for bandlimited inputs. Figure 3.60 shows a block diagram of the basic V/UV/S analysis algorithm. For wideband inputs, the pre-processing stage consists of a 200Hz hpf to remove dc, hum and low frequency noise. For telephone line inputs a second order inverse filter is used to normalise the effects of varying telephone lines.

For wideband inputs, the five parameters considered were:

(i) Energy in the signal,
(ii) Zero-crossing rate of the signal,
(iii) Autocorrelation coefficient at unit sample delay,
(iv) First predictor coefficient, and
(v) Energy of the prediction error.

These measurements were shown to provide a high degree of separability between the three classes of signal for wideband inputs. However, for telephone quality inputs, the bandlimiting of the telephone line considerably reduced the effectiveness of all the parameters in separating the classes of voiced speech, unvoiced speech and silence. For example, the absence of signal energy above 3kHz significantly reduced the number of zero crossings for unvoiced speech.

To find an effective set of parameters that would be capable of reliably distinguishing between the three signal classes for telephone line inputs, the 70 parameters were studied. Using a set of training data, the probability density functions for each of the
parameters were estimated. Those parameters that provided little or no separation between voiced, unvoiced speech and silence were eliminated from consideration. The remaining 36 parameters were studied as to their effectiveness in classifying telephone line inputs. A knockout type optimisation was used to obtain the five most effective parameters for classifying signals according to an error-weighting scheme. Several combinations of different test sets of data and error weights were investigated. The final step in the analysis method of figure 3.61 is a distance computation to determine whether a test signal is voiced, unvoiced or silence.

In conclusion, it was shown that, depending upon the weight attached to the various types of mis-classification, a set of optimal features can be found that minimises the weighted mis-classification error rate. For telephone line inputs, the results showed that reliable discrimination between voiced and voiceless sounds (silence plus unvoiced speech) can be attained at error rates fairly close to those obtained with wideband input signals.

3.6.1.1 V/UV/S based on Linear Delta Modulation and Zero-Crossing Rate

C K Un and H H Lee (ref 73) have proposed a method of V/UV/S discrimination of speech based on the results of counting bit alternations of the bit stream from linear delta modulation (LDM) of the speech signal and zero crossings (ZC) of a bandpass filtered output of the decoded LDM signal. The block diagram of the V/UV/S discriminator is shown in figure 3.62(a) and the diagram of the bit alternation and zero crossing rate measurement system is shown in figure 3.62(b).
Once the bit alternation and zero crossing rates of the LDM and BPF outputs are determined, the two measurements $Z_L(nT)$ and $Z_B(nT)$ are used to classify the input speech signal as either silence, unvoiced or voiced speech. The decision logic is of a simple synchronous sequential scheme. Based on the states of $Z_L(nT)$ and $Z_B(nT)$ that belong to one of the three (low, middle and high) regions and the decision state from the previous V/UV/S measurement, the decision logic decides whether the input speech is voiced (V), unvoiced (UV), or silence (S) at each sampling instant. Some delay (up to 10mS) is also used to smooth any decision errors.

Once the optimum LDM sampling frequency and bandpass filter parameters had been determined, the optimum threshold levels for the bit alteration rate and the zero-crossing rates of the corresponding LDM and bandpass filtered signals needed to be found. The silence segment yielded the highest zero-crossing rate while the voiced speech segment exhibited the lowest rate. Hence the proper threshold values by which a ternary decision can be made was set by obtaining average values of bit alteration and zero crossing rates for the three classes of speech.

To test the proposed method, computer simulations with male and female speech bandlimited to 3.4kHz, were performed. The total length of the test sentences was about 25s. The tests were recorded in two different environments, that is, in a soundproof room and in a computer room. When the input speech was noise free, the error rate of the speech discriminator was said to be 3 per cent. When the speech had background noise, the error rate increased to 6 per cent, mostly occurring at the boundaries of signal classes. The
system had not been tested with telephone-quality speech. It is however unclear whether the detector can successfully discriminate bandlimited fricative sounds from background noise.

3.6.2 Wideband Speech Analysis

Discerning unvoiced speech from background noise is difficult when the signal is filtered to 3.4kHz. If the input signal is allowed a wider bandwidth before being analysed then detectors can become more simplified and operate more reliably. S G Knorr(ref 74) has proposed a system which is capable of providing an accurate indication of whether or not a given speech segment is voiced or unvoiced. The approach is based upon the well-known technique of spectral energy distribution of voiced and unvoiced speech, whereby the basic V/UV decision is obtained by a simultaneous comparison of the short time spectral energy distribution of a lowpass filtered (voiced) and a highpass filtered (unvoiced) waveform.

The basic V/UV decision system is shown in figure 3.63. All high-frequency components above $f_c = 1000$Hz are removed by a 6-pole lowpass filter for the voiced decision channel. The upper frequency limit of $f_c = 1000$Hz assures that at least the first formant frequency of voiced speech is contained at the output of the filter and essentially all the unvoiced speech components are removed in the voiced decision channel. A 6-pole highpass filter removes all frequency components below $f_c = 5000$Hz for the unvoiced decision channel. The frequency band between 1000Hz to 5000Hz is not used so as to minimise ambiguities in the decision.
The short time spectral integration is performed by precision full-wave rectification and a peak detection scheme. The short-time integrations of both voiced and unvoiced spectral energies are determined by choosing decay time constants which are:

\[ \tau_{\text{dec}}(V) = 5.5 \text{mS} \quad \text{and} \quad \tau_{\text{dec}}(UV) = 1 \text{mS}. \]

These parameters had been experimentally determined. The system also requires proper weighting of the relative spectral energies for voiced and unvoiced speech which is accomplished by a V/UV decision mix control setting, \( d_{\text{uv}} \).

To overcome the problem of the sensitivity to burst of unvoiced speech components during voiced segments in the pitch initiation phase, a delay of about 4mS is included in the unit to defer an unvoiced decision every time speech goes from 'V' to 'UV'. Only for unvoiced segments exceeding the time frame of about 4mS is interpreted as unvoiced. The transition from 'UV' to 'V' speech is achieved in 2mS or more as a delay of 2mS is incorporated in the detector when switching from UV to V. During silent intervals of speech, the decision may be either voiced or unvoiced which will, to a large extent, depend upon the signal processing prior to the V/UV decision process. If an automatic gain control circuit (agc) is employed which has a very fast attack time of about 3mS, but a very long release time of 2s, the silent intervals between words is usually interpreted as voiced.
The system was tested for speech with and without moderate amount of background noise. Male, female and child speakers were used for the performance tests where the worst mis-classification rate was found to be < 0.6 per cent. The V/UV decision mix was originally set for a male's voice and no readjustment was necessary thereafter to accommodate for the other speakers.

3.6.3 V/UV Using Context

The simpler methods of voicing decisions are usually made in a context-free manner on successive portions of speech. The voicing decision is improved by the use of context, \(^{\text{ref 75}}\) in fact an improvement was found using just the previous segment of speech. For each statistic, instead of looking for a threshold that selects voiced segments, two thresholds are used, one if the last segment was called voiced and another if the last segment was unvoiced. A typical improvement obtained by allowing this 'hysteresis' in the voicing decision is said to be a 15 per cent drop in the error rate.

The parameter of low-frequency-energy is used as an example to demonstrate the improvement in voicing decision. It is shown that during the onset of voicing, the low frequency energy tends to rise abruptly but trails off slowly during the offset. By eye it was determined that when voicing commences, the energy rises to within 17dB of its overall maximum value within 100mS (one segment), while voicing ceases for all practical purposes when the low frequency
energy falls to 23dB below its maximum in 100ms. Therefore a two-threshold voicing criterion are set:

(a) If the last segment was called voiced, call the current segment unvoiced if the low frequency energy is 23dB below the maximum, otherwise it is voiced;

(b) If the last segment was called unvoiced, call the current segment voiced if the low frequency energy rises to within 17dB of the maximum, otherwise it is unvoiced.

Using the criterion, this produces voiced and unvoiced energy distributions that are now measured relative to whichever threshold current. These distributions have a smaller area of ambiguity (overlap) than the distributions measured without using hysteresis hence the error rate can be reduced. It is considered that the use of two-sided context and the use of more distant speech segments could reduce the V/UV classification error rate even further.

The description in this section regarding the classification of V/UV/S speech was intended to be simple and brief. For a more exhaustive review on this subject the reader is directed to reference 72 where it will be discovered that many more sophisticated techniques exist (such as pattern recognition approaches etc).

This section concludes our review of the state-of-the-art of relevant research which now forms a basis for the remainder of the topics to be discussed in this thesis.
Chapter 4

Voiced-Unvoiced Excitation Classification

4.1 Introduction

Some of the numerous processing techniques that have been developed to extract the voiced/unvoiced (V/UV) dichotomy of speech, just discussed in section 3.6, can be used in the study of bandwidth compression and quality enhancement. We have to some degree exploited these current techniques that are available for the V/UV classification. In this chapter, we briefly re-examine some of the techniques used, together with the necessary modifications to tailor these methods to the requirements of the bandwidth compression and quality enhancement systems discussed in chapters 5 and 6.

The parameters considered for speech classification for our purposes are:

1. Waveform inspection.
2. Zero crossing (axis crossing) of the time waveform.
4. Short time energy level of the speech signal.
5. The first sample shift autocorrelation value.

Before discussing these topics separately, we must first review the experimental methods used to perform our investigations.
4.2 **Experimental Procedure**

The experimental procedures adopted for the topics discussed in this chapter are common to the practice of all of the investigations used in this study. This description is intended to be brief as the actual methods used are very likely to become superseded by the technological advances of the University's Computer Centre in the foreseeable future, therefore it is unlikely that future research workers will utilise methods used here.

An outline of the practical method adopted is illustrated by the chart in figure 4.1. The steps are itemised below:

a) The speech was recorded onto analogue tape using a tape recorder (Ferrograph Logic 7 series) and microphone (condenser type). The speech duration recorded was about 5-10 seconds.

b) The speech data was then converted into binary form by an analogue-to-digital converter interface under the control of a Hewlett Packard Computer, HP2100. It was at this point that the sampling frequency $F_s$ was set. The speech data was then stored for processing on 9-track digital magnetic tape.

c) To enable the University's main computer to read this data, two conversion routines were performed; one to convert the binary digits to ASCII (American Standard Code for Information Interchange) characters. A list of the ASCII code is included in Table 4.1 for the reader's interest. The second tape
transcription routine involved the addition of all of the ICL 1900 protocols (i.e. file and block markers etc) such that the 7-track magnetic tape could be used on an ICL mag-tape drive peripheral to provide the source data for processing.

d) A large bulk of the data processing for the development and simulation of algorithms was achieved by the University's main computer, the ICL 1900S. The source programs were stored on filing-system files (on-disk), the data was read into the program core from the magnetic tapes produced by steps (a) to (c), and the control of the job to run the program was performed by submitting a deck of cards to be batch processed.

The output from the job-run is produced on printouts from the line-printer in the form of a listing of the program and numerical output (SNR's etc). There was also the facility of obtaining 'sketch' plots on the line-printer paper itself which was useful in the program development stage. If the job-run was successful then higher quality graph plots were provided together with a tape or filing-system file containing the processed data.

e) The output data file existed in the same form as the input file. Two transcription programs were now run (one by the ICL 1900 computer and the other using the HP2100). The first removed the ICL protocols and the second converted the ASCII characters into binary digital form.
f) The data was now changed back to analogue form using a digital-to-analogue converter peripheral device on the HP2100 computer. The sampling frequency was again set to the input value $F_s$ (or a modified sampling frequency, depending upon the simulation algorithm).

g) The data was also recorded for future reference by the analogue tape recorded as in (a). The speech is played back via the tape recorder's built in audio amplifier to an external high quality loud speaker.

The test words used in our experiments were: "sister, father, S K Harvey, shift, thick, fist, talk, spent, vote", spoken by a male speaker. The signal was filtered to 300-7600Hz by a 6th-order Butterworth filter (Bar & Stroud model No. EF2) and sampled at 16kHz. The analogue to digital conversion used linear PCM to 10-bits/sample accuracy.

Due to the limited computing facilities available when the research work was undertaken, we needed to permit ourselves to assume that this small specimen of data is "typical" speech. We acknowledge that all of the conclusions discussed in this thesis are based upon results from data using a single speaker.

4.3 Waveform Inspection

One of the initial methods used for speech classification was afforded by straight forward inspection of the time waveform of the speech signal. This is in fact using the human brain as the computer linked to the human eye as the peripheral device. The
waveform inspection was performed by noting various characteristics related to the different speech sounds. Voiced signals were noted for the relatively large amplitudes and quasi-periodic structure while unvoiced signals had shown smaller amplitudes and possessed a more irregular structure than voiced signals. The waveforms associated with /s/ sounds generally appeared to have a higher amplitude range than other unvoiced signals, also they seemed to have a relatively long duration i.e. about 60-100mS. The /s/ signal appeared also to be characterised by a regular and consistently high zero-crossing rate throughout its time waveform. The stop consonants seemed on average to produce waveforms of shorter duration, i.e. about 40mS, with an initial step and decaying excursion. The higher rate of zero-crossings appeared generally confined to the trailing portion of this unvoiced waveform. The waveforms of /f/ and /θ/ were usually the lowest in amplitude and appeared to have a more irregular rate of zero crossings. The /ʃ/ signal possessed a waveform rather similar to the /s/ signal but had a lower rate of zero-crossings. The consonant sounds that are less frequently used in spoken English language (ref 3) such as /dz/ and /tʃ/ have not been investigated in this study.

A whole series of time waveforms relating to unvoiced signals are shown in figure 4.2. These words are taken from the 10 words used in experiments. The x-axis is divided into blocks of 512 samples (i.e. 32mS duration); the y-axis shows the relative amplitude level of the signal with ±2000.0 being the maximum range of the signal. It may be seen from longer waveforms of the speech that the start and finish of the unvoiced segments used may be discerned to within, say, 4mS for stops and 10mS for fricative consonants.
The corresponding plots for the spectral cross-sections of some of these unvoiced sounds are shown in figure 4.3. The spectra of the utterances shown are for /t/, /s/, /k/, /v/, /ʃ/ and /f/ taken from our 10-word sequence. The spectra were each obtained using 512 points of speech data, weighted by a Hanning window (see section 6.4.1). The 512-sample blocks were also augmented with 512 zero-valued samples before the DFT was applied. The dB-scale is normalised with respect to the total energy contained within each block.

It can be seen that the spectrum for the /t/ utterance exhibits two peaks; one at the low frequency range, associated with the initial step and decay in the time function, and the other peak between 3.5 and 5kHz which is associated with the high frequency "ringing" at the trailing edge of the /t/-waveform. The /s/ appears to have a single prominent peak centred between 4 and 5kHz with comparatively lower energy outside the peak range. The spectrum of the /k/ appears to have three significant peaks; one in the low frequency region, one at about 2kHz and one just above 4kHz. There is also a noticeable rise in the spectrum at just about 6kHz. The spectrum associated with the /v/ sound shows two peaks, one at a low frequency position due to the voicing and one at about 3kHz due to the resonance between the teeth and lips. The spectrum due to the /ʃ/ sound in 'shift' shows a broad prominent peak just below 3kHz and a secondary peak in the spectral envelope at about 5.5kHz. The spectrum associated with the /f/ sound has an almost flat trend with a slight roll-off towards the high frequency region. It also has a random fine structure which is not surprising since perceptually the /f/ sound is probably the most noise-like speech sound in the English language.
Some further time and spectral plots were taken of a few more unvoiced sounds, these are /z/, /g/, /b/, /d/ and /g/ in figures 4.4 and 4.5 respectively. Firstly, the time plots of the /z/ and /g/ utterances appear very similar to the /s/ time waveform. However their spectra show significant differences. The /z/ spectrum has an envelope peak between about 4 and 5kHz as does the /s/ spectrum but this peak is accompanied by some voicing energy at about 200Hz. This voiced energy is not easily discerned from the time waveform of the /z/ signal. The /g/ spectrum is similar to the /z/ spectrum but both the low frequency and the high frequency peaks are broader than for the /z/.

The waveform of the /b/ utterance exhibits a sharp leading edge as the other stop consonants. There is also some higher frequency ringing after the initial step. The spectrum of the /b/ sound appears to show a definite peak at about 400Hz and a decaying envelope thereafter. There are also noticeable rises in the spectrum at about 1.6 and 2.2kHz.

The time waveform of the /d/ utterance has an initial step lasting for a shorter duration than the /b/ sound but the high frequency content is far more noticeable which is also shown in its respective spectrum by the broad peak centred at about 4.6kHz. The initial step and decaying time function is also manifested in the spectrum by the low frequency peak at about 500Hz with a decaying envelope and regular fine structure beyond the 500Hz region. The fine structure becomes more random-like after 4kHz.
Finally, the waveform of the /g/ shows a double step and decaying function. The second step is seen to be more prominent than the first. The associated spectrum could well be confused with that of a vowel as it shows three peaks with an approximately fine structure. The 'harmonics' are possibly due to the spacing of the two steps in the time waveform. The spectral peaks are perhaps due to the ringing after the step functions.

Only a handful of the unvoiced utterances of the 23 consonants (ref 3) from the English language have been briefly examined. So from this standpoint, we can observe that unvoiced speech exhibits time and spectral characteristics that are as equally significant as those of voiced speech.

4.4 Special Distribution Method for V/UV Switching

We have already discussed the notion of separating voiced speech from unvoiced speech using a spectral distribution criterion in section 3.6.2. In this section, we wish to verify and perhaps modify the method by Knorr (ref 74) to suit some of the schemes used for bandwidth compression later on in this study.

We start by referring back to figure 3.63 and modelling the scheme by software (indicated in section 4.2) as specified in the reference. The exception to the original specifications was the
highpass filter used for the unvoiced decision channel. The lowpass and highpass filters were implemented as Butterworth filters of 6-pole and 15-pole respectively. The program listings for the Butterworth filters are included in the appendix A4. The voiced/unvoiced decision mix, $d_{uv}$, is a parameter that will require optimisation, or at least experimental investigation. The full-wave rectification was achieved using the 'ABS' function from the computer library and the short time integration was set up using the 1st order discrete lowpass filter expression:

$$H(z) = \frac{G}{1 - a z^{-1}}$$

(4.1)

where $G$ is the normalizing factor, incorporated to set the dc gain to unity

i.e. $H(z) \bigg|_{z=1} = \frac{G}{1 - a} = 1$

\[ 
\therefore G = 1 - a \tag{4.2} 
\]

The integrator takes the form of figure 4.6.
The value 'a' determines the time constant of the integrator and takes the value

\[ a = \exp(- T/\tau) \]

where \( T = \frac{1}{F_s} \) is the sampling interval

and \( \tau \) is the time constant.

For the prescribed values used by Knorr (ref 74), \( \tau(v) = 5.5 \text{mS} \) for the voiced-path and \( \tau(uv) = 1 \text{mS} \) for the unvoiced-path.

The remainder of the elements in the system of figure 3.63 i.e. the comparator and delayed switch are considered straightforward in operation and are not explained here.

During silent pauses of the speech signal, the decision may be either voiced or unvoiced. An automatic gain control (agc) circuit (ref 74) may be employed having a very fast attack time of about 3mS, and a very long release time of about 2S. By using this kind of agc prior to the V/UV detector, the silent intervals between words were usually interpreted as voiced speech. The agc element was implemented as a switched first order filter arrangement as shown in figure 4.7. The scheme is such that if the magnitude of the current output sample \( |y_1| \), of the network is greater than the previous gain value \( G_{i-1} \), then the time constant is set to 3mS. If the condition
does not apply then the time constant is set to 2 seconds. If $x_i$ is
the $i$th input sample and $y_i$ is the corresponding output sample then

$$y_i = G_i x_i$$ \hspace{1cm} (4.3)

where $G_i$ is the current gain factor.

$G_i$ is determined as follows

Let $a_1 = \exp(- T_1/\tau_1)$
and $a_2 = \exp(- T_1/\tau_2)$
where $\tau_1 = 3\text{mS}$ and $\tau_2 = 2\text{s}$
Also for unity gain at d.c.

$$G_1 = 1 - a_1$$
and $$G_2 = 1 - a_2$$

so $G_i = g \mid y_i \mid + a G_{i-1}$ \hspace{1cm} (4.4)

where $a = a_1$ and $g = g_1$ if $\mid y_i \mid > G_{i-1}$
or $a = a_2$ and $g = g_2$ if $\mid y_i \mid \leq G_{i-1}$

The result of using this agc network of figure 4.7 is shown in
figure 4.8 for the signal silence-/es/ from "S K Harvey". Figure
4.8(a) shows the input signal, $x_i$. Figure 4.8(b) shows the gain
function, $g_i$, and figure 4.8(c) shows the output signal, $y_i$. It can
be seen that at the onset of the voiced /c/, the gain function uses
the $3\text{mS}$ rise time constant and subsequently changes to the $2\text{s}$
decay time constant. For sustained voiced sounds, the gain is almost constant and high. For prolonged silence intervals, the gain is comparatively low and constant.

With the agc network preceding the voiced/unvoiced detector shown in figure 3.63, all that is left to be determined is the V/UV decision mix, $d_{uv}$. A suitable starting value for $d_{uv}$ was selected and the output plot of the two short time spectral integrators were compared. These two quantities were fed to the input of the comparator in figure 3.63, therefore the output of the comparator was controlled by $d_{uv}$.

The results of the short time spectral integrators are shown in figure 4.9 and 4.10 respectively. These are taken for the input signal "silence-/es/" shown in figure 4.8(a). By increasing $d_{uv}$, the magnitude of the signal in figure 4.10 increases with respect to that in figure 4.9, i.e. the measurement of unvoiced energy increases with respect to the voiced-energy measurement.

The system shown in figure 3.63 was operated with the filters having a passband gain of 0dB and the integrators and agc network having a unity gain at dc. The system was tested using the 10 word sequence "sister, father, S K Harvey, shift, thick, fist, talk, spent, vote", of which the signal 'silence-/es/' from "S K Harvey" yielded the output signal shown in figure 4.11. It was found that when $d_{uv}$ was set to 17.5, all of the fricative sounds were separated from the speech along with small amounts of silent-interval noise. Also,
some of the stop consonants were clipped. Increasing the value of $d_{uv}$ seemed to prevent the clipping of the stop consonants, but at the expense of including extra silent-interval noise.

The optimisation of the system of figure 3.63 was not continued at this stage as it was desired to use the method with different filter bandwidths for the voiced and unvoiced filters. This was such that the method can be employed in conjunction with one of the bandwidth compression algorithms, discussed in Chapter 6 of this thesis. For this latter objective, the system was then used to compare the spectral energy distribution of the 300 to 7600Hz speech signal. The comparison was made between the spectral energy within the 300 to 3400Hz band and the 3000-6000Hz frequency band. This was achieved by setting the voiced and unvoiced filters to these respective cut-off frequencies. With the voiced/unvoiced adjustment $d_{uv}$ set to 0.5, the response of the short time spectral integrators is shown in figures 4.12 and 4.13 for the input 'silence-/es/' . The output of this V/UV switch is shown in figure 4.14 which indicates that it is the /s/ signal only that is selected as 'unvoiced' with the silence-interval noise being interpreted as 'voiced'. For the whole of the 10-word sequence used as the input data, the V/UV switch was found to favour /s/ utterances only as unvoiced speech, with the parameters of the V/UV switch set to the values stated, all other unvoiced speech was selected as voiced speech. The reason for this is possibly due to the fact that the /s/ signal has the most prominent energy in the 3000 to 6000Hz range and a comparatively small amount in the 300 to 3400Hz range which is apparent from the /s/ spectrum in figure 4.3(b). The V/UV system
operating in this mode was found very useful for some of the bandwidth compression algorithms to be described, however if it is required to select the /s/ utterances from the 300 to 7600Hz speech then the method described so far may not be the most efficient since it requires filtering and this is costly in terms of computer execution time. For physical realisations, however, the process can be achieved by straightforward analogue techniques.

4.5 Detection based upon Zero-Crossing Rate of the time Waveform

An alternative method for only discerning /s/ utterances relies upon the zero-crossing rate of the input signal. By inspecting the waveform of figure 4.2(b) of the 300-7600Hz /s/-signal, it can be seen that the waveform has a consistently higher axis crossing (zero-crossing) rate than the waveforms of the other speech utterances. For this reason it would appear a relatively simple task to sift the /s/ utterances from the 300 to 7600Hz signal. The scheme, operating upon the waveform zero-crossing rate, measures the time duration, $t_z$, between successive zero-crossings of the original signal (which is sampled at 16kHz). The algorithm is such that when $t_z$ falls below 0.25mS and is maintained below this value for at least 4mS or more, then an /s/ was deemed to be present at the input. The switch is again reset to the 'non-/s/' position if, after an /s/ has been detected, $t_z$ rises above the 0.25mS threshold for at least 2mS. The maximum delay of switching is thus 4mS. The zero-crossing strategy facilitates switching in a non block-synchronous manner i.e. the switch can run on a continuous mode.
A flow chart illustrating the operation of the zero-crossing duration detector is shown in figure 4.15. Initially, the sample counter NCNT is set to zero, also are the 'lead-in' and the hangover counters L and H. The thresholds are also preset, i.e. the duration between adjacent zero-crossings $T_c$ is set to 0.25mS; the lead-in threshold $L$ is set to $4/T$ and the hangover threshold $H$ is set to $2/T_m$S. The scheme then enters the main loop by accepting a sample, $x_i$, from the input speech data which is then buffered before being applied to the detector itself. The length of the buffer is equivalent to the maximum delay taken by the detector which for our case is 4mS. The samples at the input to the buffer (i.e. $x_i$) are also passed through a hard-limiter to determine their sign. If the sign of the current sample and the previous sample are unchanged then the sample counter is incremented. If the sign does change then this indicates that a zero-crossing in the input sequence is encountered. The time between the current zero-crossing and the previous one is noted i.e.

$$t_z = (NCNT)T \quad (4.5)$$

where

$$T = 1/F_s$$

The sample counter is then reset, i.e. NCNT=0, and the value $t_z$ is compared with the threshold $T_c$. If the duration is less than the preset value then the 'lead-in' counter, $Zl$, is incremented. When the lead-in counter exceeds the preset lead-in time i.e. $Zl$ is greater or equal to $L$, then the output samples from the buffer, $y_i$, is accepted as a sample from an /s/ utterance. Once this decision is made, the hangover counter is also set to the duration of the hangover, i.e. equivalent to 2mS.
When the value of $t_z$ becomes greater than $T_c$, then the subsequent steps after this decision are by-passed and the hangover counter is decremented.

If the hangover counter is greater than zero then the output from the buffer is still taken as an /s/-sample as the hangover period will not have expired. If the hangover period does expire and the value of $t_z$ is above the threshold then the lead-in counter will also be reset (i.e. $L = 0$) such that when the next /s/ utterance is detected, the lead-in will need to be accumulated again. The algorithm returns to the input sample stage until all the data is processed.

Initially the parameters of the detection scheme were determined by referring to the plots of $t_z$ versus time distribution for the utterance /sIs/ in "sister". Part of this distribution is shown in figure 4.16(a) and (b). It may be seen that for the /s/ utterance, $t_z$ remains below 0.2mS except for the occasional rise to 0.3mS i.e. for a shorter time than the hangover period of 2mS, whereas the value of $t_z$ is around 1.0mS or above for the voiced /I/. The lead-in and hangover periods are included such that the detector follows the trend of $t_z$ and is unaffected by the narrow holes and peaks in the distribution of $t_z$. The value of 0.25mS was therefore selected for the threshold $T_c$. The distribution of $t_z$ versus time was also checked for other unvoiced utterances as well as voiced speech and silence-intervals (not shown) and it was found that the value of $t_z$ consistently remained below the threshold, $T_c$, only for /s/ utterances. The output of the ZC-detector is shown in figure 4.16(d).
The detection scheme operating with the above parameters was tested by our 10-word sequence "sister, father, S K Harvey, shift, thick, fist, talk, spent, vote". It was found that only the five /s/'s were selected with all other sounds including silent interval microphone noise being rejected. When the result of the detection scheme was compared with the results of selecting the /s/ sounds by visual inspection of the speech waveform it was found that the zero-crossing detection scheme added up to 100 samples (≈6ms) of silence noise before the leading edge and after the trailing edge of the /s/ utterances from the data used. This result was considered quite acceptable for the applications discussed later.

The zero-crossing detection scheme was then tested for selecting /s/ signals from 300 to 3400Hz speech. This bandlimited speech signal is used to simulate speech that has undergone transmission via the telephone system. Since at the very outset we are only concerned by enhancing bandlimited speech signals, all other degrading factors effected by the telephone network transmission (outlined in chapter 2) are not considered. The means of bandlimiting the signal was achieved by using a 255 point finite impulse response filter (FIR) which is described in appendix A 1.2. The resultant input and output signal for the /Is/ in "sister" is shown in figure 4.17(a) & (b). It can be seen that the voiced /I/ is almost unaffected by the bandlimiting process but the unvoiced /s/ is heavily attenuated and distorted; this phenomenon will be discussed further in Chapter 5.

When the zero-crossing detector was applied to the bandlimited speech, the distribution of $t_z$ versus time plot appeared as shown in figure 4.18. It illustrates that the value of $t_z$ varies in a much
more random fashion for the /s/ fricative and from this it would be
more difficult to discern the boundaries and duration of an /s/
utterance and silence pauses than by using the plot of $t_z$ shown in
figure 4.16 (i.e. by using the knowledge of the 300 to 7600Hz
wideband signal). If the plot of figure 4.17 is compared to the
distribution of $t_z$ versus time for bandlimited silent interval noise
as shown in figure 4.19 then it should be apparent that the
zero-crossing detector would have a more difficult and unwieldy task
in separating the bandlimited /s/ from silent interval noise as
indeed was the case. It is difficult to decide where the threshold
$T_c$ should lie on the plots of figures 4.18 and 4.19. If multiple
thresholds and hysteresis were to be used however, then the
performance of the detector might be improved. This notion was not
pursued here.

The zero-crossing detection scheme was also checked to ascertain
whether other unvoiced sounds could be separated from 300 to 7600Hz
continuous speech. Figure 4.20 shows the variation of $t_z$ versus
time for the fricatives /ʃ/ from "shift" and /f/ from "fist". The
plot of $t_z$ for silent-interval noise is also shown. It can again be
seen that as with the bandlimited /s/ case, the variation of $t_z$
corresponding to the /f/ sound does not have a consistently low
value as it did for the wideband /s/ signal. It was therefore
concluded that the separation of silence and /f/ sounds from all
other signals by means of a single threshold would again be
difficult and unreliable. The discrimination of the /f/ sound
appears more encouraging although more work is required in
identifying this sound from silence noise. Lowpass filtering the
$t_z(t)$ function does seem likely to prove helpful.
4.6 Detection based upon Power Level Measurements

A third parameter that is commonly used to discern unvoiced speech sounds is to measure the short term energy level of the signal. With the use of one or more thresholds, the input signal can be classified for further processing or recognition purposes. In our tests, we have found it possible to select /s/ sounds from the data used in the experiments. Also we find this method particularly suitable for selecting the /f/ and /θ/ fricative sounds from the speech data. This would be directly applicable to our speech quality enhancement experiments described in Chapter 5.

The scheme considered is shown in figure 4.21. The algorithm is a mean square detector where the mean is taken over 8mS (128 samples). The mean square signal, \( y_i = \langle x_i^2 \rangle \), is given by

\[
y_j = \frac{1}{N_1} \sum_{i=1}^{N_1} x_{j-i}^2
\]

where \( N_1 = 8mS/T = 8 F_s \)

The mean square signal \( y_i \) is compared with two thresholds, \( T_1 \) and \( T_2 \). \( T_1 \) is set by the peak value of the input signal, \( x_i \) and \( T_2 \) is a fixed threshold. If a condition arises such that

\[
T_2 < y_j < T_1
\]

then the output switch, \( S_1 \), closes then giving a positive detection of the desired signal.
The reason for using two thresholds is that the variable threshold, $T_1$, provides an indication of the sound level of the talker. The determination of $T_1$ is achieved by passing the input signal $x_i$ through the agc circuit used in section 4.4 and shown in figure 4.7. With $y_i$ as the input to the agc, the threshold becomes dependent upon the previous maximum power level of the speech. The agc network has a decay time constant of 2 seconds such that if the talker changes his speech volume or is substituted by another person, then the threshold adapts to the new speaker level.

The threshold, $T_1$, is set by reducing the agc output by $1/20^{th}$ of its value. This was a tentative value ascertained by noting the range of values of $y_i$ for the highest level of voiced speech and comparing this range with that for the two /f/ sounds and /θ/ sound from the 10-word data used in the experiment. The mean square value of the /ft/ in "shift" is shown in figure 4.22. It can be seen that the mean square value for the /f/ and the trailing edge of the /t/ are almost the same, for this reason 10mS hysteresis was applied to the comparator in figure 4.21 such that if a logical 0 or 1 emerges from the comparator for less than 10mS then it is ignored. This had the effect of rejecting the /t/ utterance.

The threshold $T_2$ was fixed and is used as a 'floor' value. If $y_j$ falls below this value then the input signal $x_i$ is considered to be silent-interval noise. For our experiments, $T_2$ was set to a value of 10.0. The floor threshold need not be adaptive as it can be preset to a value determined by the audio equipment used (which generates the spurious noise) and it not necessary to set $T_2$ to be dependent upon the level at which the talker speaks.
When the system was run with the parameters as stated, it did indeed successfully select the two /f/ utterances and /θ/ sound from the 10-word sequence but a small amount of spurious noise was also accepted by the detector. A criticism of the method is that the calculation of the mean of the power level by a sliding window method was computationally large in execution time. One method of overcoming this was to use an infinitely long duration non-uniform window weighting function i.e. that facilitated by a first order lowpass filter instead of the sliding window. The idea now is modified to that as shown in figure 4.23.

In this method, the output of the envelope detector provides an indication of the short time power level of the input speech. The time constant of the first order filter was set to 10mS so as to smooth the fluctuations in the signal level during a fricative sound. The smoothed power level for a stretch of silent-interval noise was computed during a speech pause of about 0.25S duration. The speech pause was selected by visual inspection of the input speech waveform. This "dc" level was then subtracted from the short term power level of the input speech so as to compensate for the audio equipment noise.

The peak level sensor network used to discern the talker's level of voiced speech was the same agc network shown in figure 4.7. Again, a hysteresis of 10mS was used to differentiate between the trailing edges of some stops from the short term power level of the fricative sounds. The short term power level of an /f/ sound in "father" with the dc (power level) subtracted is shown in figure 4.24.
When this system was tested for the 10-word sequence, the two /f/ and one /θ/ utterances were selected but again some spurious signal was also included. If the fixed threshold, \( T_2 \), was increased so as to remove the spurious noise detection then it was found that the wanted signals were clipped at the start and finish.

The detection method was again modified in an attempt to reject the spurious noise by cascading the power level detector method with the zero-crossing duration method as shown in figure 4.25.

The measurement of the short term power level comparison is by the same method as that shown in figure 4.23 and the zero-crossing method, ZC, is the same as the flow chart of figure 4.15. The parameters and thresholds of the power level comparator were set to the values, as those in figure 4.23, whilst the values for the zero-crossing detection scheme were re-adjusted from those previously set for selecting the /s/ utterances. The value of the threshold \( T_c \) was set to 0.5mS and the lead-in and hangover period, \( L \) and \( H \), were both set to 2mS. Figure 4.26 shows the plot for the zero-crossing duration versus time for the silence-/f/ signal from the beginning of the word "father". Having \( T_c = 0.5 \text{mS} \) did show that there was some "separability" between the /f/ sound and silence noise although the reliability of detection was increased by the combination of this zero-crossing duration measurement with the power level measurement system. The number of false detections from the 10-word sequence was 10 with the zero-crossing detection and 5 from the power level detection method but only 3 false detections were found when the two methods were combined.
4.7 Detection of Voiced Speech based upon the First Shift Autocorrelation Coefficient

The first-shift normalised correlation function \( \rho_1 \), for a block of \( x_i \) samples of signal data is given by

\[
\rho_1 = \frac{\sum_{i=1}^{N_1-1} x_i x_{i+1}}{\sum_{i=1}^{N_1} x_i^2} \quad (4.7)
\]

The value of \( \rho_1 \) will range between \( \pm 1 \) depending upon whether the sequence \( \{x_i\} \) is fully correlated or negatively correlated. The value of \( \rho_1 \) was calculated for concatenating blocks of 8mS (i.e. \( N_1 = 128 \)) of speech samples from the 10-word input sequence. It was found that for the 16kHz sampled signal the value of \( \rho_1 \) was around 0.9 to 0.95 for voiced speech and about 0.5 or less for the non-voiced speech (where non-voiced speech comprises of unvoiced speech and silence). Occasionally, the value of \( \rho_1 \) increased to 0.7 during the non-voiced speech signals. The value of \( \rho_1 \) fell to between -0.3 and -0.5 however for the /s/ utterances. With the distribution of \( \rho_1 \) as stated, it would appear a relatively straightforward task to select voiced speech from non-voiced speech; when the method was tested using a single threshold of 0.89, it was found possible to select all of the voiced sounds from the 10-word sequence.

There was the problem with the voiced switch prematurely clipping some of the trailing portions of the voiced sounds as it was discovered that this was due to the value of \( \rho_1 \) falling to about 0.7
during the last 8mS block of these voiced segments. The problem of clipping was overcome by incorporating the value of $P_1$ for the previous 8mS block into the strategy. The decision switch now operates every 16mS (comprising of two 8mS blocks), instead of every 8mS. The decision is now that if for the $j^{th}$ and $(j-1)^{th}$ block

\[ P_1(j-1) > T_s \quad \text{and} \quad P_1(j) > T_s \]  

(4.8)

where $T_s$ is the system threshold parameter, voiced speech is deemed to be present for both blocks $(j-1)$ and $j$. Should the speech correlation be below the threshold for two consecutive block viz:

\[ P_1(j-1) < T_s \quad \text{and} \quad P_1(j) < T_s \]  

(4.9)

we infer that the speech in blocks $j$ and $(j-1)$ is voiceless. For the cases when

\[ P_1(j-1) < T_s \quad \text{and} \quad P_1(j) > T_s \]

and

\[ P_1(j-1) > T_s \quad \text{and} \quad P_1(j) < T_s \]  

(4.10)

both blocks $j$ and $(j-1)$ assume the same decision as blocks $(j-2)$ and $(j-3)$ i.e. the decision for the current 16mS of speech is identical to the decision made during the preceding 16mS block.

The results of this scheme are shown in figure 4.27, figure 4.27(a) shows the original utterance for /s-ka/ from "S K Harvey" and figure 4.27(b) shows the response of the voiced-switch. The result clearly
shows that the vowel /a/ is separated from the remaining non-voiced speech in the 16 blocks of signal specimen shown. Table 4.2 shows the values of $P_1$ for the blocks of speech. Each block is of 16mS duration, therefore according to our strategy there are two values of $P_1$ calculated for each block and both values must be above the threshold for the decision to change from non-voiced to voiced. It can be seen from table 4.2 that the value of $P_1$ falls below the threshold in block 173 which from figure 4.27(a) shows that the signal at this instant should be voiced. However, according to equation (4.10), the decision was indeed maintained as "voiced" as required.

Table 4.2 also indicates that although for unvoiced utterances, the value of $P_1$ ranges between -0.3 and 0.78 for voiceless fricatives, this value increases to 0.83 during intervals of silence-noise. The values of $P_1$ for the remaining signal in the 10 word sequence were typified by the values shown in table 4.2 which influenced our choice for the value of the threshold, i.e. $T_s = 0.89$.

The results described for the voice-switch are based upon experiments using 300 to 7600Hz input speech. When the input was replaced by speech bandlimited from 300 to 3400Hz it was found that the operation of the voice-switch was unaffected. With the decision strategy and parameters unchanged, it was apparent that the segments of 300 to 3400Hz speech selected as "voiced" were the same as those segments from the 300 to 7600Hz bandlimited speech. One obvious difference in performance of the voice switch is that it tended to switch to the "voiced-decision" about 8-16mS earlier when operating
<table>
<thead>
<tr>
<th>Block No.</th>
<th>$\rho_1$ 300-7600Hz</th>
<th>$\rho_1$ 300-3400Hz</th>
<th>Speech Sound (c.f. fig. 4.27)</th>
</tr>
</thead>
<tbody>
<tr>
<td>161</td>
<td>-0.26</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>161</td>
<td>-0.24</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>-0.26</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>162</td>
<td>-0.20</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>-0.26</td>
<td>0.75</td>
<td></td>
</tr>
<tr>
<td>163</td>
<td>-0.21</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>-0.14</td>
<td>0.76</td>
<td></td>
</tr>
<tr>
<td>164</td>
<td>0.52</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>0.39</td>
<td>0.78</td>
<td></td>
</tr>
<tr>
<td>165</td>
<td>0.45</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>0.69</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>166</td>
<td>-0.28</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>0.78</td>
<td>0.83</td>
<td></td>
</tr>
<tr>
<td>167</td>
<td>0.70</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>0.79</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>168</td>
<td>0.78</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>0.67</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>169</td>
<td>0.77</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>0.86</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>170</td>
<td>0.82</td>
<td>0.85</td>
<td></td>
</tr>
<tr>
<td>171</td>
<td>0.83</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>171</td>
<td>0.85</td>
<td>0.88</td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>0.92</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>0.91</td>
<td>0.90</td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>0.87</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>173</td>
<td>0.94</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>0.94</td>
<td>0.94</td>
<td></td>
</tr>
<tr>
<td>174</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>0.95</td>
<td>0.95</td>
<td></td>
</tr>
<tr>
<td>175</td>
<td>0.96</td>
<td>0.96</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>0.92</td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>176</td>
<td>0.97</td>
<td>0.97</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.2. Comparison of the first-lag autocorrelation coefficient for successive blocks of wideband and band-limited speech.
upon the 300 to 3400Hz speech than when applied to the 300-7600Hz speech. It was also true that the switch changed back to the "non-voiced" decision at about 8-16mS later when operating upon the input speech with the smaller bandwidth. This is because the lowpass filtering action tends to smooth the speech data with a consequential increase in the value of $p_1$. Column 3 of table 4.2 shows the corresponding values of $p_1$ for the 300Hz to 3400Hz speech which illustrates this point. The most notable example of this smoothing phenomenon is the fact that $p_1$ is around 0.78 in blocks 160-164 for the bandlimited signal, whereas in column 2 this value was around -0.29 for the 300 to 7600Hz signal. Referring again to figure 4.27(a) we note that blocks 160-164 contain speech samples of an /s/ utterance from the utterance "S K Harvey". Even with this in mind, we find that the switching strategy is sufficiently biased in favour of the "voiced" decision to maintain a reliable operation for the bandlimited 10 word data specimen used in our experiment.

4.8 Discussion

In this chapter, we have discussed various means of identifying certain speech signals from continuous speech based upon the properties of these speech signals. The methods described are by no means considered as a thorough investigation, both in terms of the number of schemes considered and of the depth in which each technique was pursued; rather, the notions used have been intended to be an exploratory investigation into the front-end requirement of some of the bandwidth compression and bandwidth enhancement
algorithms elucidated in the next two chapters. It is again emphasised that all the experiments are based upon the 10-word sequence of data described in section 4.1. It therefore is very likely that if these methods for voiced and unvoiced speech classification were used for trials on speech from other sources (e.g. female and child speech) then the parameters would almost certainly require re-adjustment.

From our experiments using the schemes described, we find that the tasks of selecting /s/-sounds from 300 to 7600Hz speech and selecting voiced-speech from 300 to 7600 speech were by far the simplest and most reliable to achieve. The detection of voiced speech from the 300 to 3400Hz bandlimited speech proved to be a similar effort to that when the wider bandwidth speech was used.

Apart from the visual inspection of the waveform, the selection of other specific speech sounds i.e. stops and fricatives, resulted in more difficult and unreliable procedures. However, at least the experimentation provided an insight into some of the problems encountered, e.g. separating /f/ and /θ/ sounds from silent interval equipment noise.

Having now discussed some techniques involved in classifying speech signals, we now turn our attention to bandwidth enhancement schemes which is the topic of our next chapter.
Fig. 2.1. Schematic diagram of the human vocal mechanism (ref. 3)

Fig. 2.2. Schematic vocal tract profiles for the production of English vowels. (Adapted from Potter, Kopp and Green)
Table 2.1. *Vowels*

<table>
<thead>
<tr>
<th>Degree of constriction</th>
<th>Tongue hump position</th>
<th>front</th>
<th>central</th>
<th>back</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>/i/      eve</td>
<td>/i/' bird</td>
<td>/u/ boot</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/i/      it</td>
<td>/o/' over (unstressed)</td>
<td>/u/ foot</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>/e/      hate*</td>
<td>/æ/ up</td>
<td>/o/ obey*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>/e/      met</td>
<td>/o/ ado (unstressed)</td>
<td>/o/ all</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>/æ/      at'</td>
<td>/o/</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* These two sounds usually exist as diphthongs in GA dialect. They are included in the vowel table because they form the nuclei of related diphthongs. (ref 3)

Table 2.2. *Fricative consonants*

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Voiced</th>
<th>Voiceless</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labio-dental</td>
<td>/v/  vote</td>
<td>/f/  for</td>
</tr>
<tr>
<td>Dental</td>
<td>/θ/  then</td>
<td>/θ/  thin</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/z/  zoo</td>
<td>/s/  see</td>
</tr>
<tr>
<td>Palatal</td>
<td>/ʒ/  azure</td>
<td>/ʃ/  she</td>
</tr>
<tr>
<td>Glottal</td>
<td>/h/  he</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.3. *Stop consonants*

<table>
<thead>
<tr>
<th>Place of articulation</th>
<th>Voiced</th>
<th>Voiceless</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>/b/  be</td>
<td>/p/  pay</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/d/  day</td>
<td>/t/  to</td>
</tr>
<tr>
<td>Palatal/velar</td>
<td>/g/  go</td>
<td>/k/  key</td>
</tr>
</tbody>
</table>
Fig. 2.3 Articulatory profiles for the English stop consonants. (After Potter, Kopp and Green)

Table 2.4. Nasals (ref. 3)

<table>
<thead>
<tr>
<th>Place</th>
<th>NasalSound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labial</td>
<td>/m/ me</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/n/ no</td>
</tr>
<tr>
<td>Palatal/velar</td>
<td>/ŋ/ sing</td>
</tr>
<tr>
<td></td>
<td>(no initial form)</td>
</tr>
</tbody>
</table>

Fig. 2.4 Vocal profiles for the nasal consonants. (After Potter, Kopp and Green)
Table 2.5. *Glides and semi-vowels* (ref. 3)

<table>
<thead>
<tr>
<th>Place</th>
<th>Sound</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palatal</td>
<td>/j/</td>
<td>you</td>
</tr>
<tr>
<td>Labial</td>
<td>/w/</td>
<td>we</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(no final form)</td>
</tr>
<tr>
<td>Palatal</td>
<td>/rs/</td>
<td>read</td>
</tr>
<tr>
<td>Alveolar</td>
<td>/l/</td>
<td>let</td>
</tr>
</tbody>
</table>

Fig. 2.5 Vocal tract configurations for the beginning positions of the glides and semivowels. (After Potter, Kopp and Green)
Fig. 2.6 Schematic diagram of the human ear showing outer, middle and inner regions. The drawing is not to scale. For illustrative purposes the inner and middle ear structures are shown enlarged (ref. 3).

Fig. 2.7 Simplified diagram of the cochlea uncoiled (ref. 3).

Fig. 2.8 Schematic cross section of the cochlear canal. (Adapted from Davis, 1957)
<table>
<thead>
<tr>
<th>Noise</th>
<th>dB</th>
<th>Relative energy</th>
<th>Sound pressure dyne/cm²</th>
<th>Typical examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deafening</td>
<td>120</td>
<td>$1,000,000,000,000$</td>
<td>200</td>
<td>Threshold of pain</td>
</tr>
<tr>
<td></td>
<td>110</td>
<td>$100,000,000,000$</td>
<td></td>
<td>Thunder</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Gunfire</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pneumatic drill</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Steam whistle</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Large machine shop</td>
</tr>
<tr>
<td>Very loud</td>
<td>90</td>
<td>$1,000,000,000$</td>
<td></td>
<td>Underground railway</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Busy street</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Noisy factory</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Inside aeroplane</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Loud public address system</td>
</tr>
<tr>
<td></td>
<td>80</td>
<td>$100,000,000$</td>
<td>2</td>
<td>Noisy office</td>
</tr>
<tr>
<td>Loud</td>
<td>70</td>
<td>$10,000,000$</td>
<td></td>
<td>Suburban train</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Typewriters</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Radio set—full volume</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average factory</td>
</tr>
<tr>
<td>Moderate</td>
<td>60</td>
<td>$1,000,000$</td>
<td>0.2</td>
<td>Large shop</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average office</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Quiet motor car</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Quiet office</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Average house</td>
</tr>
<tr>
<td>Faint</td>
<td>50</td>
<td>$100,000$</td>
<td></td>
<td>Public library</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Country road</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>$10,000$</td>
<td>0.02</td>
<td>Quiet conversation</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Rustle of paper</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Whisper</td>
</tr>
<tr>
<td>Faint</td>
<td>30</td>
<td>$1,000$</td>
<td></td>
<td>Quiet church</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Still night in the country</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>100</td>
<td>0.002</td>
<td>Sound-proof room</td>
</tr>
<tr>
<td>Very faint</td>
<td>10</td>
<td>10</td>
<td></td>
<td>Threshold of hearing</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>0.0002</td>
<td></td>
</tr>
</tbody>
</table>
Fig 2.9(a)

Time waveform of speech for a 153 mS segment corresponding to the utterance "S.K. Harvey".
Fig 2.9(b) Voiced "a" segment.

Fig 2.9(c) Unvoiced "s" segment.
FIG. 2-10  LONG TIME AVERAGED SPECTRAL DENSITY OF SPEECH (REF 7)
Fig 2.11(a) Short-time spectrum for an "a" sound.

Fig 2.11(b) Short-time spectrum for an "u" sound.
FIG. 2.12 (a) COMPLETE CIRCUIT OF BT TELEPHONE No. 706

FIG. 2.12 (b) SPEECH WITHOUT REGULATOR SHOWING OPERATING OF ANTI-SIDETONE INDUCTION COIL (REF 8)
FIG. 2.12 (c) Continued. SIMPLIFIED DIAGRAM SHOWING OPERATION AT REGULATOR (REF 8)

FIG. 2.13 STONE BRIDGE FEED CIRCUIT (REF 8)
Fig 2.14 The trunk network (ref.8)
(a) CIRCUIT WITH 2/4 WIRE TRANSITIONS

(b) DETAILS OF HYBRID TRANSFORMER

FIG. 2.15 2/4—WIRE TRANSITION (REF 8)
Fig. 2.16 ORATOR Acoustic Unit (ref.12)
FIG. 2.17 SIMPLIFIED BLOCK DIAGRAM OF TELEPHONE BALANCING UNIT (REF 16)
FIG 3.1 THE INPUT-OUTPUT CHARACTERISTIC OF A TIME-IN Variant QUANTIZER (REF 19)
FIG. 3.2 ADAPTIVE QUANTIZER CHARACTERISTICS FOR (a) LOW SIGNAL POWER AND (b) LARGE SIGNAL POWER (REF. 7)

\[ e_i' = e_i + q_i \]

FIG. 3.3 THE DPCM CODEC (REF. 19)
Fig. 3.4 a and b. Delta modulator with single integration (ref. 3)

Fig. 3.5 Waveforms for a delta modulator with single integration (ref. 3)

Fig. 3.6 Adative delta modulator with single integration (ref. 3)
Fig. 3.7 1st order constant factor d.m. encoder. (After Jayant ref. 25)

FIG. 3.9 BLOCK DIAGRAM OF THE CVSD (REF 27)
Fig 3.10  Source-system representation of speech production (ref. 3)

Fig 3.11  Block diagram of the original spectrum channel vocoder. (After Dudley, 1939b)

Fig 3.12  Spectrogram of speech transmitted by a 15-channel vocoder (ref. 3)
Table 3.1 Consonant intelligibility for a vocoder. Percent of initial consonants heard correctly in syllables (logatoms). (After Halsey and Swaffield)

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Intelligibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>b</td>
<td>90%</td>
</tr>
<tr>
<td>l</td>
<td>97%</td>
</tr>
<tr>
<td>r</td>
<td>90%</td>
</tr>
<tr>
<td>w</td>
<td>90%</td>
</tr>
<tr>
<td>f</td>
<td>74</td>
</tr>
<tr>
<td>m</td>
<td>85</td>
</tr>
<tr>
<td>s</td>
<td>94</td>
</tr>
<tr>
<td>sh</td>
<td>90</td>
</tr>
<tr>
<td>h</td>
<td>100</td>
</tr>
<tr>
<td>n</td>
<td>99</td>
</tr>
<tr>
<td>t</td>
<td>91</td>
</tr>
<tr>
<td>th</td>
<td>43</td>
</tr>
<tr>
<td>k</td>
<td>85</td>
</tr>
<tr>
<td>p</td>
<td>77</td>
</tr>
<tr>
<td>v</td>
<td>96</td>
</tr>
<tr>
<td>none</td>
<td>70</td>
</tr>
</tbody>
</table>

Fig 3.13 Phonetic pattern-matching vocoder. (After Dudley, 1958)

Fig 3.14 Autocorrelation vocoder. (After Schroeder, 1959, 1962)
Fig 3.15 Parallel-connected formant vocoder. (After Munson and Montgomery)

Fig 3.16 Cascade-connected formant vocoder. (After Flanagan and House)

Fig 3.17 Block diagram of a computer-simulated speech synthesizer. (After Flanagan, Coker and Bird)
Fig 3.18 Method for describing and synthesizing the short-time speech spectrum in terms of Fourier coefficients. (After Pirogov)

Fig 3.19 Model of speech production. (ref. 30)
Fig 3.20  a and b. Analysis and synthesis operations for the homomorphic vocoder. (After Oppenheim)

Fig 3.21 Synthesis method for the maximum likelihood vocoder. Samples of voiced and voiceless excitation are supplied to a recursive digital filter of $p$-th order. Digital-to-analog ($D/A$) conversion produces the analog output. (After Itakura and Saito, 1968)
Fig 3.22 Approximations to the speech spectrum envelope as a function of the number of poles of the recursive digital filter. The top curve, $|X(j\omega)|$, is the measured short-time spectral density for the vowel /a/ produced by a man at a fundamental frequency 140 cps. The lower curves show the approximations to the spectral envelope for $p = 6, 8, 10$ and 12.
(After ITAKURA and SAITO, 1970)

Fig 3.23 Synthesis from a recursive digital filter employing optimum linear prediction.
(After ATAL and HANAUER, 1971)
Fig 3.24 Block diagram of the Vobanc frequency division-multiplication system.  
(After Bogert, 1956)

Fig 3.25 The Hilbert Transformer.  \( h(t) \) and \( H(j\omega) \) are Fourier transform pairs.
Fig 3.26 Diagram for computer simulation of the analytic rooter. 
(After Schröder, Flanagan and Lundby)

Table 3.2 Eighth-order Butterworth filter cutoff frequencies in cps (ref. 3)

<table>
<thead>
<tr>
<th>Channel</th>
<th>BPF 1</th>
<th>BPF 1/2</th>
<th>Formants nominally in passband</th>
</tr>
</thead>
<tbody>
<tr>
<td>Channel 1</td>
<td>238–714</td>
<td>119–357</td>
<td>F 1</td>
</tr>
<tr>
<td>Channel 2</td>
<td>714–1428</td>
<td>357–714</td>
<td>F 1 or F 2</td>
</tr>
<tr>
<td>Channel 3</td>
<td>1428–2142</td>
<td>714–1071</td>
<td>F 2 or F 3</td>
</tr>
<tr>
<td>Channel 4</td>
<td>2142–2856</td>
<td>1071–1428</td>
<td>F 3</td>
</tr>
</tbody>
</table>
FIG. 3·27  FILTERING OF A SPEECH SIGNAL BY CONTIGUOUS BAND-PASS FILTERS
FIG 3.28 SPEECH SYNTHESIS FROM SHORT-TIME AMPLITUDE AND PHASE DERIVATIVE SPECTRA

Fig 3.29 Programmed analysis operations for the phase vocoder. (After FLANAGAN and GOLDEN)
**Fig 3.30** Voice-Excited Vocoder (ref.48)

**Fig 3.31** SPECTRAL FLATTENER AND SPEECH SYNTHESIZER FOR 700 CPS BASE BAND (ref.48)
Fig 3.32 Four-band encoder using low-pass translation and APCM encoding in each band. (ref. 49)

Fig 3.33 Integer-band sampling technique for digital encoding of speech sub-bands. (ref. 49)
FIG. 3.34 (a) Quadrature mirror filter bank. (b) A spectral interpretation. (REF. 44)
Fig 3.35 (a) Relative comparison of quality of 16-kb/s sub-band coding against ADPCM coding (based on listener preference) for different ADPCM coder bit rates. (b) Relative comparison of quality of 9.6-kb/s sub-band coding against ADM coding for different ADM coder bit rates. (ref. 49)
Fig. 3.36 Transform coding scheme. (ref. 53)

Fig. 3.37 Interpretation of bit assignment rule. (ref. 7)

Fig. 3.38 Representation of side information as equal spaced samples of the smoothed DCT spectrum. (ref. 7)
Fig 3.39 Components of the speech spectrum model, (a) formant structure, (b) pitch structure, and (c) combined model. (ref. 7)

Fig 3.40 Illustration of operation of the "vocoder-driven" ATC algorithm, (a) DCT spectrum and spectral estimate, (b) bit assignment and (c) reconstructed signal. (ref. 7)
Fig 3.41 Block diagram of a generalized predictive coder with two stages of prediction: a predictor $P_s$ based on the short-time spectral envelope and another predictor $P_d$ based on the pitch periodicity. The filter $F$ is chosen to reduce the perceived distortion of quantizing noise. (ref. 7)

Fig 3.42 An example showing the envelope of the output noise spectrum (dotted curve) shaped to reduce perceived distortion ($F$ as in Eq.3.85) and the corresponding speech spectrum (solid curve). (ref. 7)
Fig 3.43 Qualitative model for frequency division of voiced sounds. (ref. 46)

**Fig. 3.44** General systems configuration for combining TDHC and waveform coding (ref. 60)
Amplitude

---

Complex zeros

---

4 Complex zeros

---

**Fig. 3.45** SHAPE DESCRIPTORS REPRESENTING A SEGMENT OF SPEECH WAVEFORM

---

**Fig. 3.46** Spectrum of speech coding transmission rates (nonlinear scale) and associated quality. (ref. 7)
<table>
<thead>
<tr>
<th>CLASS</th>
<th>EXAMPLES</th>
<th>BIT RATE kbit/s</th>
<th>QUALITY</th>
<th>COMPLEXITY</th>
<th>CURRENT STATUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Waveform</td>
<td>PCM</td>
<td>64</td>
<td>excellent</td>
<td>low</td>
<td>widespread use</td>
</tr>
<tr>
<td></td>
<td>DPCM</td>
<td>32, 16</td>
<td>very good</td>
<td></td>
<td>demo, available</td>
</tr>
<tr>
<td></td>
<td>(with companding or level adaptation)</td>
<td>8</td>
<td>poor</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Channel vocoder</td>
<td>2.4</td>
<td>fair</td>
<td>high</td>
<td>widespread use</td>
</tr>
<tr>
<td></td>
<td>LPC vocoder</td>
<td>2.4</td>
<td>fair</td>
<td>high</td>
<td>some use, increasing</td>
</tr>
<tr>
<td></td>
<td>Formant vocoder</td>
<td>1.2</td>
<td>can be good</td>
<td>high</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>Pattern-matching vocoder</td>
<td>0.8</td>
<td>poor</td>
<td>very</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>Phoneic vocoder</td>
<td>0.1 - 0.2</td>
<td>poor</td>
<td>very</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>Baseband vocoder</td>
<td>9.6</td>
<td>good</td>
<td>fairly high</td>
<td>demo, available</td>
</tr>
<tr>
<td></td>
<td>APC</td>
<td>16</td>
<td>very good</td>
<td>fairly high</td>
<td>designs</td>
</tr>
<tr>
<td></td>
<td>or RELP</td>
<td>8, 4.8</td>
<td>good</td>
<td>high</td>
<td>available - research</td>
</tr>
<tr>
<td></td>
<td>Sub-band coding</td>
<td>16</td>
<td>good</td>
<td>medium</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>SBC with harmonic compression</td>
<td>8</td>
<td>good</td>
<td>high</td>
<td>research</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Adaptive transform coding</td>
<td>16, 8</td>
<td>very good</td>
<td>high</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>ATC with harmonic compression</td>
<td>8, 4.8</td>
<td>very good</td>
<td>high</td>
<td>research</td>
</tr>
<tr>
<td></td>
<td>Time-encoded speech</td>
<td>4-16</td>
<td>not yet proved</td>
<td>medium</td>
<td>research</td>
</tr>
</tbody>
</table>

Table 3.4 Summary of the properties of speech coding systems. (ref.17)
FIG. 3.48 INPUT-OUTPUT VOLTAGE CHARACTERISTIC OF NETWORK FOR SIGN MULTIPLICATION (REF 47)

FIG. 3.49 VOCODER SYSTEM FOR TRANSMITTING HIGH-QUALITY SPEECH OVER A NARROW BANDWIDTH. ALLOCATION OF FREQUENCIES IN THE TRANSMISSION PASS-BAND IS SHOWN IN FIG. 3.50 (REF 47)
Fig 3.50 Allocation of frequencies in the pass-band of the transmission medium. The nine vocoder channels occupy the frequency band from 330 c/s to 780 c/s in an inverted order. They represent the speech information in the bands 2.0 - 2.3 - 2.6 - 3.0 - 3.5 - 4.2 - 5.2 - 6.5 - 8.1 - 10.0 kc/s. The uncoded speech band from 80 c/s to 2 kc/s is shifted upward to the frequency band from 780 c/s to 2.7 kc/s. (ref. 47)
FIG. 3-51 PILOT INJECTION CONTROLLED OVERTONE REPRODUCTION SYSTEM (PICOR (ref 63))
Fig 3.52 Two-band coding scheme. The lower band coded with 4-bit ADPCM. The upper band is coded with 3-bit/sample APCM for 56 kbit/s transmission or 4-bit/sample APCM for 64 kbit/s. (ref. 68)

FIG. 3-53 CHARACTERISTIC OF THE HIGH FREQUENCY BOOSTER USED FOR HIGH FREQUENCY ENHANCEMENT.
Fig. 3.54 (a) Baseband spectrum
(b) 3-band spectral folding
(c) 3-band spectral translation. (ref. 70)

Fig. 3.55 Receiver for baseband coder that uses integer-band spectral folding. (ref. 70)

Fig. 3.56 Receiver for baseband coder that uses integer-band spectral translation. (ref. 70)

Fig. 3.57 The filter $H(z)$ for $F_w = 3F_B$ (ref. 70)
Fig. 3.58 Low and high frequency regeneration applied to telephone bandwidth speech (ref 71)
(a) SYLLABIC EXPANDER/NOISE-GATE (REF 69)

(b) TYPICAL TRANSFER FUNCTIONS OF THE EXPANDER/NOISE-GATE

FIG. 3·59
Fig 3.60 — Block diagram of silence-unvoiced-voiced classification system. (ref. 72)

Fig 3.61 — Flow chart of knockout optimization algorithm. (ref. 72)

Fig 3.62 (a) Block diagram of voiced/unvoiced/silence discriminator. (ref. 73)
Fig 3.62 (b) Block diagram of bit alternation and zero crossing rates measuring system. (ref. 73)

Fig 3.63 Block diagram of the basic V/UV decision system. (ref. 74)
FIG. 4.1  EXPERIMENTAL PROCEDURE
<table>
<thead>
<tr>
<th>Character</th>
<th>7-Bit Hexadecimal Number</th>
<th>Character</th>
<th>7-Bit Hexadecimal Number</th>
<th>Character</th>
<th>7-Bit Hexadecimal Number</th>
<th>Character</th>
<th>7-Bit Hexadecimal Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>NUL</td>
<td>00</td>
<td>SP</td>
<td>20</td>
<td>@</td>
<td>40</td>
<td>\</td>
<td>60</td>
</tr>
<tr>
<td>SOH</td>
<td>01</td>
<td>!</td>
<td>21</td>
<td>A</td>
<td>41</td>
<td>a</td>
<td>61</td>
</tr>
<tr>
<td>STX</td>
<td>02</td>
<td>&quot;</td>
<td>22</td>
<td>B</td>
<td>42</td>
<td>b</td>
<td>62</td>
</tr>
<tr>
<td>ETX</td>
<td>03</td>
<td>#</td>
<td>23</td>
<td>C</td>
<td>43</td>
<td>c</td>
<td>63</td>
</tr>
<tr>
<td>EOT</td>
<td>04</td>
<td>$</td>
<td>24</td>
<td>D</td>
<td>44</td>
<td>d</td>
<td>64</td>
</tr>
<tr>
<td>ENQ</td>
<td>05</td>
<td>%</td>
<td>25</td>
<td>E</td>
<td>45</td>
<td>e</td>
<td>65</td>
</tr>
<tr>
<td>ACK</td>
<td>06</td>
<td>&amp;</td>
<td>26</td>
<td>F</td>
<td>46</td>
<td>f</td>
<td>66</td>
</tr>
<tr>
<td>BEL</td>
<td>07</td>
<td>'</td>
<td>27</td>
<td>G</td>
<td>47</td>
<td>g</td>
<td>67</td>
</tr>
<tr>
<td>BS</td>
<td>08</td>
<td>(</td>
<td>28</td>
<td>H</td>
<td>48</td>
<td>h</td>
<td>68</td>
</tr>
<tr>
<td>HT</td>
<td>09</td>
<td>)</td>
<td>29</td>
<td>I</td>
<td>49</td>
<td>i</td>
<td>69</td>
</tr>
<tr>
<td>LF</td>
<td>0A</td>
<td>*</td>
<td>2A</td>
<td>J</td>
<td>4A</td>
<td>j</td>
<td>6A</td>
</tr>
<tr>
<td>VT</td>
<td>0B</td>
<td>+</td>
<td>2B</td>
<td>K</td>
<td>4B</td>
<td>k</td>
<td>6B</td>
</tr>
<tr>
<td>FF</td>
<td>0C</td>
<td>.</td>
<td>2C</td>
<td>L</td>
<td>4C</td>
<td>l</td>
<td>6C</td>
</tr>
<tr>
<td>CR</td>
<td>0D</td>
<td>-</td>
<td>2D</td>
<td>M</td>
<td>4D</td>
<td>m</td>
<td>6D</td>
</tr>
<tr>
<td>SO</td>
<td>0E</td>
<td>,</td>
<td>2E</td>
<td>N</td>
<td>4E</td>
<td>n</td>
<td>6E</td>
</tr>
<tr>
<td>SI</td>
<td>0F</td>
<td>/</td>
<td>2F</td>
<td>O</td>
<td>4F</td>
<td>o</td>
<td>6F</td>
</tr>
<tr>
<td>DLE</td>
<td>10</td>
<td>0</td>
<td>30</td>
<td>P</td>
<td>50</td>
<td>p</td>
<td>70</td>
</tr>
<tr>
<td>DC1</td>
<td>11</td>
<td>1</td>
<td>31</td>
<td>Q</td>
<td>51</td>
<td>q</td>
<td>71</td>
</tr>
<tr>
<td>DC2</td>
<td>12</td>
<td>2</td>
<td>32</td>
<td>R</td>
<td>52</td>
<td>r</td>
<td>72</td>
</tr>
<tr>
<td>DC3</td>
<td>13</td>
<td>3</td>
<td>33</td>
<td>S</td>
<td>53</td>
<td>s</td>
<td>73</td>
</tr>
<tr>
<td>DC4</td>
<td>14</td>
<td>4</td>
<td>34</td>
<td>T</td>
<td>54</td>
<td>t</td>
<td>74</td>
</tr>
<tr>
<td>NAK</td>
<td>15</td>
<td>5</td>
<td>35</td>
<td>U</td>
<td>55</td>
<td>u</td>
<td>75</td>
</tr>
<tr>
<td>SYN</td>
<td>16</td>
<td>6</td>
<td>36</td>
<td>V</td>
<td>56</td>
<td>v</td>
<td>76</td>
</tr>
<tr>
<td>ETB</td>
<td>17</td>
<td>7</td>
<td>37</td>
<td>W</td>
<td>57</td>
<td>w</td>
<td>77</td>
</tr>
<tr>
<td>CAN</td>
<td>18</td>
<td>8</td>
<td>38</td>
<td>X</td>
<td>58</td>
<td>x</td>
<td>78</td>
</tr>
<tr>
<td>EM</td>
<td>19</td>
<td>9</td>
<td>39</td>
<td>Y</td>
<td>59</td>
<td>y</td>
<td>79</td>
</tr>
<tr>
<td>SUB</td>
<td>1A</td>
<td>:</td>
<td>3A</td>
<td>Z</td>
<td>5A</td>
<td>z</td>
<td>7A</td>
</tr>
<tr>
<td>ESC</td>
<td>1B</td>
<td>;</td>
<td>3B</td>
<td>(</td>
<td>5B</td>
<td>\</td>
<td>7B</td>
</tr>
<tr>
<td>FS</td>
<td>1C</td>
<td>&lt;</td>
<td>3C</td>
<td>)</td>
<td>5C</td>
<td>)</td>
<td>7C</td>
</tr>
<tr>
<td>GS</td>
<td>1D</td>
<td>=</td>
<td>3D</td>
<td>\</td>
<td>5D</td>
<td>ALT</td>
<td>7D</td>
</tr>
<tr>
<td>RS</td>
<td>1E</td>
<td>&gt;</td>
<td>3E</td>
<td>I</td>
<td>5E</td>
<td>ESC</td>
<td>7E</td>
</tr>
<tr>
<td>US</td>
<td>1F</td>
<td>?</td>
<td>3F</td>
<td>-</td>
<td>5F</td>
<td>DEL, RUBOUT</td>
<td>7F</td>
</tr>
</tbody>
</table>
Fig 4.2(a) /t/ in sister.

Fig 4.2(b) /s/ in "S.K. Harvey".
Fig 4.2(c) /k/ in "S.K. Harvey".

Fig 4.2(d) /v/ in "Harvey".
Fig 4.2(e) \( f(t) \) in "shift".

Fig 4.2(f) \( f(t) \) in "fist".
Fig 4.3(a) Power spectrum of /t/ from "sister".

Fig 4.3(b) Power spectrum of /s/ from "S.K. Harvey".
Fig 4.3(c) Power spectrum of /k/ from "S.K. Harvey".

Fig 4.3(d) Power spectrum of /v/ from "Harvey".
Fig 4.3(e) Power spectrum of /f/ from "shift"

Fig 4.3(f) Power spectrum of /f/ from "flat"
Fig 4.4(a) /s/ in "zoo".

Fig 4.4(b) /s/ in "azure".
Fig 4.4(c) /b/ in "be".

Fig 4.4(d) /d/ in "day".
Fig 4.5(a) Power spectrum of /z/ from "zoo".

Fig 4.5(b) Power spectrum of /z/ from "azure".
Fig. 4.5(c) Power spectrum of /b/ from "be".

Fig. 4.5(d) Power spectrum of /d/ from "day".
FIG. 4:6 FIRST ORDER INTEGRATOR

FIG. 4:7 AGC NETWORK
(a) Input waveform to agc network of fig. 4.7.

(b) Gain function applied to the amplifier in fig. 4.7.

Fig 4.8
Fig 4.8(c) Output from the age network in fig 4.7.
Fig 4.9  Output from the low-pass spectral integrator in fig 3.63.

Fig 4.10  Output from the high-pass spectral integrator in fig 3.63.

(d_{uv} = 17.5)
Fig 4.11 Output from the V/UV decision system in fig 3.63.
Fig 4.12  Output from low-pass spectral integrator. The filter bandwidth is now set to 0 - 3000Hz.

Fig 4.13  Output from high-pass spectral integrator. The filter bandwidth is now set to 3-6 kHz, $d_{uv} = 0.5$. 
Fig 4.14 Output from V/UV decision system of fig 3.63 with the new filter bandwidths.
Set threshold, $T_c = 0.25\text{ms}$
Lead-in, $L = 4\text{ms}/T$
Hangover, $H = 2\text{ms}/T$

Set counts $NCNT, Z1, Z2 = 0$

Input sample, $X$
Buffer $X$
Output $Y$  \( X \) buffered by $L$ samples

$S_i = \text{Sign}(X)$

is $S_i = S_i - 1$ ?
Yes \( \text{NCNT} = \text{NCNT} + 1 \)
No

Time = $\text{NCNT} \times T$  \( (t_a) \)

$\text{NCNT} = 0$  \( \sim \text{Reset duration counter} \)

is $\text{Time} < T_c$ ?
NO

$Z1 = Z1 + 1$  \( \sim \text{Increment lead-in counter} \)

$Z1 \geq L$ ?
NO

$Y$ is /s/ sample
$Z2 = H + 1$  \( \sim \text{Set hangover} \)

$Z2 = Z2 - 1$  \( \sim \text{Decrement hangover counter} \)

$Z2 \geq 0$ ?
NO

$Y$ is /s/ sample
end of data ?

$T = 1/F_s$

$\text{NCNT} = \text{NCNT} - 1$ \( \text{1----} \)

Time = $\text{NCNT} \times T$

$Z1 = Z1 + 1$  \( \sim \text{Increment lead-in counter} \)

$Z1 \geq L$ ?
NO

$Y$ is /s/ sample

$Y$ is not /s/ \( \sim \text{Reset lead-in counter} \)

$Z2 = Z2 - 1$  \( \sim \text{Decrement hangover counter} \)

$Z2 \geq 0$ ?
NO

$Y$ is /s/ sample

end of data ?

Stop

FIG 4.15 FLOW CHART OF /s/ DETECTOR USING ZERO CROSSING RATE MEASUREMENTS
\( t_z (\mu s) \)

duration between successive zero crossing of the time waveform

Fig 4.16(a) Waveform zero crossing duration, \( t_z \), for an /s/ sound of 300 to 7600Hz bandwidth.

Fig 4.16(b) Waveform zero crossing duration, \( t_z \), for an /l/ sound of 300 to 7600Hz bandwidth.
(c) Input waveform to the V/UV decision scheme in fig 4.15.

(d) Output from V/UV detector based upon zero crossing rate of waveform.

Fig 4.16
Fig 4.17(a) Time waveform for an /ls/ sound, bandlimited from 300 to 7600Hz.

Fig 4.17(b) Time waveform for an /ls/ sound, bandlimited from 300 to 3600Hz.
Fig 4.18 Waveform zero crossing duration, $t_z$, for an /ls/ sound bandlimited from 300 to 3400Hz.

Fig 4.19 Waveform zero crossing duration of silent interval noise bandlimited between 300 to 3400Hz.
Fig 4.20 Waveform zero crossing duration for
(a) /f/ in "fist" and (b) /ʃ/ in "shift".
The waveforms are bandlimited from 300 to 7600Hz.
FIG. 4.21  V/UV DETECTION BASED UPON THE MEAN POWER LEVEL OF THE INCIDENT SIGNAL \( \{x_i\} \)
Fig 4.22 Mean squared value of an /f/ and a/t/ sound from "shift" obtained from the network in fig 4.21.
FIG. 4.23  MODIFIED VERSION OF DETECTOR IN FIG. 4.21
Fig 4.24 Short-time power level measurement of an /f/ sound from "father". The signal is bandlimited from 300 to 7600Hz.

(a) Input

(b) Output from the network shown in fig 4.23

\[ \text{[NB the dc power level is subtracted]} \]
FIG. 4.25  COMBINATION OF SHORT TIME POWER LEVEL MEASUREMENT AND ZERO-CROSSING MEASUREMENTS FOR V/UV DETECTION
Fig 4.26(a) Silence interval and an /t/ sound bandlimited from 300 to 7600Hz.

Fig 4.26(b) Zero crossing duration for the waveform shown in fig 4.26(a).
Fig 4.27 Input and output waveforms of /s/ - /ka/
sounds from the V/UV detector based upon
the first-lag autocorrelation coefficient.

(a) Input

(b) Output

NB The output signal is delayed by 128 samples (8mS) to synchronise
it with the decision from the analysis buffer.