1

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Acquisitions and Bibliographic Services

Direction des acquisitions et des services bibliographiques

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

**Voice Frame Reconstruction Methods
for CELP Speech Coders**

by

Tze-Wo Leung, B.Sc.(Eng.)

**A thesis submitted to the
Faculty of Graduate Studies and Research
in partial fulfillment of the requirements
for the degree of
Master of Engineering**

**Ottawa-Carleton Institute for Electrical Engineering,
Faculty of Engineering,
Department of Systems and Computer Engineering,
Carleton University,
Ottawa, Ontario, Canada**

September 15, 1993

© Tze-Wo Leung, 1993

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file / Votre référence

Our file / Notre référence

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-89912-3

Canada

Name

Tze - Wo Leung

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

Voice Frame Reconstruction Methods for

SUBJECT TERM

CELP Speech Coders

0544

SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523

Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language 0679
General 0289
Ancient 0290
Linguistics 0291
Modern 0291
Literature 0401
General 0294
Classical 0295
Comparative 0297
Medieval 0298
Modern 0316
African 0591
American 0305
Asian 0352
Canadian (English) 0355
Canadian (French) 0593
English 0311
Germanic 0312
Latin American 0315
Middle Eastern 0313
Romance 0314
Slavic and East European

PHILOSOPHY, RELIGION AND

THEOLOGY

Philosophy 0422
Religion
General 0318
Biblical Studies 0321
Clergy 0319
History of 0320
Philosophy of 0322
Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
Archaeology 0324
Cultural 0326
Physical 0327
Business Administration
General 0310
Accounting 0272
Banking 0770
Management 0454
Marketing 0338
Canadian Studies 0385
Economics
General 0501
Agricultural 0503
Commerce Business 0505
Finance 0508
History 0509
Labor 0510
Theory 0511
Folklore 0358
Geography 0366
Gerontology 0351
History
General 0578

Ancient 0579
Medieval 0581
Modern 0582
Black 0328
African 0331
Asia, Australia and Oceania 0332
Canadian 0334
European 0335
Latin American 0336
Middle Eastern 0333
United States 0337
History of Science 0585
Law 0398
Political Science
General 0615
International Law and Relations 0616
Public Administration 0617
Recreation 0814
Social Work 0452
Sociology
General 0626
Criminology and Penology 0627
Demography 0938
Ethnic and Racial Studies 0631
Individual and Family Studies 0628
Industrial and Labor Relations 0629
Public and Social Welfare 0630
Social Structure and Development 0700
Theory and Methods 0344
Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
General 0473
Agronomy 0285
Animal Culture and Nutrition 0475
Animal Pathology 0476
Food Science and Technology 0359
Forestry and Wildlife 0478
Plant Culture 0479
Plant Pathology 0480
Plant Physiology 0817
Range Management 0777
Wood Technology 0746
Biology
General 0306
Anatomy 0287
Biostatistics 0308
Botany 0309
Cell 0379
Ecology 0329
Entomology 0353
Genetics 0369
Limnology 0793
Microbiology 0410
Molecular 0307
Neuroscience 0317
Oceanography 0416
Physiology 0433
Radiation 0821
Veterinary Science 0778
Zoology 0472
Biophysics
General 0786
Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996

Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
General 0566
Audiology 0300
Chemotherapy 0992
Dentistry 0567
Education 0350
Hospital Management 0769
Human Development 0758
Immunology 0982
Medicine and Surgery 0564
Mental Health 0347
Nursing 0569
Nutrition 0570
Obstetrics and Gynecology 0380
Occupational Health and Therapy 0354
Ophthalmology 0381
Pathology 0571
Pharmacology 0419
Pharmacy 0572
Physical Therapy 0382
Public Health 0573
Radiology 0574
Recreation 0575

Speech Pathology 0460
Toxicology 0383
Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry
General 0485
Agricultural 0749
Analytical 0486
Biochemistry 0487
Inorganic 0488
Nuclear 0738
Organic 0490
Pharmaceutical 0491
Physical 0494
Polymer 0495
Radiation 0754
Mathematics 0405
Physics
General 0605
Acoustics 0986
Astronomy and Astrophysics 0606
Atmospheric Science 0608
Atomic 0748
Electronics and Electricity 0607
Elementary Particles and High Energy 0798
Fluid and Plasma 0759
Molecular 0609
Nuclear 0610
Optics 0752
Radiation 0756
Solid State 0611
Statistics 0463

Applied Sciences

Applied Mechanics 0346
Computer Science 0784

Engineering
General 0537
Aerospace 0538
Agricultural 0539
Automotive 0540
Biomedical 0541
Chemical 0542
Civil 0543
Electronics and Electrical 0544
Heat and Thermodynamics 0348
Hydraulic 0545
Industrial 0546
Marine 0547
Materials Science 0794
Mechanical 0548
Metallurgy 0743
Mining 0551
Nuclear 0552
Packaging 0549
Petroleum 0765
Sanitary and Municipal System Science 0790
Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

The undersigned hereby recommend to
the Faculty of Graduate Studies and Research
acceptance of the thesis

**“Voice Frame Reconstruction Methods
for CELP Speech Coders”**

submitted by

Tze-Wo Leung, B.Sc.(Eng.)

in partial fulfillment of the requirements
for the degree of Master of Engineering

Thesis Supervisor

Chairman,

Department of Systems and Computer Engineering

Carleton University

Abstract

For voice applications in the frame relay and digital cellular radio networks, speech frames can be lost due to either traffic congestion in the packet network or deep fade conditions occurring in the mobile radio channel. The lost frames should be replaced to minimize distortion in reconstructed output speech. In this thesis, speech frame reconstruction methods suitable for use with Code Excited Linear Prediction (CELP) coders are examined. The use of Vector Linear Prediction (VLP) for parametric extrapolation and interpolation is discussed. An adaptive energy control mechanism is introduced to suppress sudden energy explosion in the reconstructed frame. Objective and subjective results at random frame error rates up to 10% are obtained. It is shown that the overall perceptual quality is improved when compared with the frame substitution approach. More importantly, there is almost no degradation in perceptual quality with a frame error rate as high as 1%.

Acknowledgements

I am indebted to Professor Samy Mahmoud and Dr. Wilf LeBlanc for their guidance and assistance throughout the course of this thesis. In particular, I would like to thank Professor Samy Mahmoud for suggesting this interesting topic and providing such an excellent research environment here in the Department. I am also grateful to Dr. Wilf LeBlanc for his advice, patience and comments. The help received from all participants of the listening tests is greatly appreciated.

The funding support of this research is made possible by a grant from the Telecommunications Research Institute of Ontario (TRIO).

Table of Contents

Abstract	ii
Acknowledgements	iii
List of Tables	vii
List of Figures	viii
List of Acronyms	x
List of Symbols	xiii
1 Introduction	1
1.1 Background: Digital Speech Coding	1
1.2 Thesis Objective.....	2
1.3 Thesis Organization	3
2 Applications of Speech Coding (Part I): Frame Relay Networks	5
2.1 Introduction.....	5
2.2 Frame Relay Network Protocol	6
2.3 Speech Coding in the Frame Relay Network.....	9
2.3.1 Code Excited Linear Prediction (CELP).....	10
2.3.1.1 Complexity Reduction and Search Strategies.....	13
2.3.1.2 Integer-CELP: A Frame Relay Speech Coder	21
2.3.2 Voice Activity Detection over Frame Relay.....	25
2.3.3 Time-Scale Modification (TSM) of CELP Coded Speech	26
2.4 Speech Frame Reconstruction in Frame Relay.....	32
2.5 Summary	33
3 Applications of Speech Coding (Part II): Digital Cellular Radio	35
3.1 Introduction.....	35
3.2 Digital Cellular Network Architecture.....	36

3.2.1	Overview of IS-54 Radio Interface	38
3.2.2	Overview of GSM Radio Interface	40
3.3	Speech Coding for Digital Cellular Communications	42
3.3.1	Vector Sum Excited Linear Prediction (VSELP)	42
3.3.2	Regular Pulse Excited Long-Term Prediction (RPE-LTP).....	50
3.3.3	Error Protection Schemes	52
3.3.3.1	Bit Error Control Methods	52
3.3.3.2	Frame Error Recovery Methods.....	56
3.4	Summary.....	58
4	Reconstruction of Missing Speech	60
4.1	Literature Review.....	60
4.1.1	Pattern Matching Method	61
4.1.2	Pitch Detection Method	64
4.1.3	Double Sided Periodic Substitution (DSPS) Method	66
4.1.4	Packet Reconstruction for ADPCM and APC	67
4.1.5	Packet Reconstruction for CELP	68
4.2	Speech Frame Extrapolation and Interpolation.....	70
4.2.1	Vector Linear Prediction (VLP).....	71
4.2.1.1	Computational Issues	76
4.2.2	CELP Frame Reconstruction	77
4.2.2.1	Frame Gain.....	78
4.2.2.2	Short-term LPC Filter	79
4.2.2.3	Adaptive Codebook	89
4.2.2.4	Stochastic Excitation Gain.....	94
4.2.2.5	Stochastic Codebook.....	96
4.2.2.6	Adaptive Energy Control	96
4.2.2.7	Multiple Lost Frames.....	97
4.3	Summary.....	97
5	Measurements of Speech Quality	99
5.1	Introduction.....	99
5.2	Objective Measures of Speech Quality.....	99
5.2.1	Parametric L_p Distances	100
5.2.2	Log-Spectral Distance (SD).....	101
5.2.3	Likelihood Ratio (LR).....	102

5.2.4	Other Objective Distortion Measures	104
5.3	Subjective Measures of Speech Quality	105
5.3.1	A-B Preference Test.....	105
5.3.2	Mean Opinion Score (MOS) Test.....	106
5.3.3	Other Subjective Measures	106
5.4	Evaluation of Speech Reconstruction Schemes.....	106
5.5	Summary	113
6	Conclusions	114
6.1	Conclusions.....	114
6.2	Suggestions for Future Research	116
6.2.1	Frame Relay Speech Coder.....	116
6.2.2	Speech Frame Reconstruction.....	117
	References	119
Appendix A	Sampling Rate Conversion	125
Appendix B	Cholesky Decomposition	133
Appendix C	Levinson-Wiggins-Robinson (LWR) Recursion	137
Appendix D	Mean Opinion Score (MOS) Test Instructions	142

List of Tables

Table 2.1 - State Table for FECN and BECN Bits	8
Table 2.2 - ICELP Frame Format	21
Table 2.3 - TSM Specification of Received CELP Frame.....	26
Table 2.4 - Time-Scale Modification Statistics	29
Table 3.1 - 8 kb/s VSELP Coder Bit Allocation.....	42
Table 3.2 - 13 kb/s RPE-LTP Bit Allocation	50
Table 3.3 - VSELP Parameter Class Bit Assignment.....	53
Table 3.4 - RPE-LTP Parameter Class Bit Assignment	54
Table 4.1 - Frame Gain LP Coefficients.....	78
Table 4.2 - VLP Reconstruction Methods for LPC Parameters.....	86
Table 5.1 - Distortion Measurements for Short-Term LPC Parameters	107
Table 5.2 - MOS Test Conditions at 0%, 1%, 2%, 5% and 10% FERs.....	109
Table 5.3 - MOS of Reconstruction Methods and Reference Conditions	110
Table 5.4 - Composite MOS of Reconstruction Methods.....	111
Table A.1 - Passband and Stopband Frequencies of Sampling Rate Converters.....	130

List of Figures

Figure 2.1 - Frame Relay Protocol Model	7
Figure 2.2 - Frame Relay Frame Structure	7
Figure 2.3 - Analysis-by-Synthesis CELP Speech Coder Structure	11
Figure 2.4 - Analysis-Synthesis Filtering of Speech	11
Figure 2.5 - Rearranged CELP Coder Structure	14
Figure 2.6 - MSVQ of CELP Stochastic Excitation	20
Figure 2.7 - ICELP Frame Buffering	20
Figure 2.8 - Frame Relay Speech Frame Structure	24
Figure 2.9 - Time-Scale Modification of CELP Coded Speech	28
Figure 2.10 - Sampling Rate Conversion	30
Figure 2.11 - Original Synthetic CELP Speech	30
Figure 2.12(a) - 4:5 Time-scale Expanded Speech	31
Figure 2.12(b) - 4:3 Time-scale Compressed Speech	31
Figure 2.12(c) - 1:2 Time-scale Expanded Speech	31
Figure 2.12(d) - 2:1 Time-scale Compressed Speech	31
Figure 2.13 - Frame Reconstruction via (a) Extrapolation and (b) Interpolation	34
Figure 3.1 - A Typical Fading Envelope at 835 MHz for a speed of 100 km/h	37
Figure 3.2 - GSM Digital Cellular Network Architecture	37
Figure 3.3 - IS-54 Time Slot and Frame Structure	39
Figure 3.4 - GSM Time Slot and Frame Structure	39

Figure 3.5 - Block Diagram of VSELP Speech Decoder.....	43
Figure 3.6 - VSELP Basis Vector Structured Codebook.....	43
Figure 3.7 - Block Diagram of RPE-LTP Speech Decoder.....	51
Figure 3.8 - VSELP and RPE-LTP Channel Coders	55
Figure 3.9 - Interleaving GSM Speech Frames	55
Figure 4.1 - Waveform Recovery via Pattern Matching Method	62
Figure 4.2 - Phase Correction in Pitch Detection Method.....	65
Figure 4.3 - Embedded Coding System.....	69
Figure 4.4 - Embedded APC Speech Coder.....	69
Figure 4.5 - Second Order LP-based Extrapolation on Frame Gain.....	80
Figure 4.6 - Second Order MSE-K-VLP-IS Extrapolation	87
Figure 4.7 - Second Order LR-LSP-VLP-IS Extrapolation	88
Figure 4.8 - Comparison of Actual and Predicted Lag Values.....	90
Figure 4.9 - Comparison of Actual and Predicted ACB Coefficients.....	91
Figure 4.10 - Lag Contour Comparison using Modified Pitch Search	93
Figure 4.11 - Stochastic Gain Extrapolation via Second Order VLP	95
Figure 4.12 - Adaptive Energy Control Mechanism.....	98
Figure 4.13 - MOS Results of Various FER and Test Conditions.....	112
Figure 4.14 - Composite MOS Results.....	112
Figure A.1 - Polyphase Filter Networks	128
Figure A.2 - Modified Polyphase Filter for $M:L$ Rate Converter	129
Figure A.3 - Frequency Response of 4:3 Sample Rate Converter	131
Figure A.4 - Frequency Response of 4:5 Sample Rate Converter	131
Figure A.5 - Frequency Response of 1:2/1:2 Sample Rate Converter.....	132

List of Acronyms

ACB	Adaptive Codebook.
APCM	Adaptive Pulse Coded Modulation.
ADPCM	Adaptive Differential Pulse Coded Modulation (ITU G.721).
AMPS	Advanced Mobile Phone System
APC	Adaptive Predictive Coding.
BECN	Backward Explicit Congestion Notification (of Frame Relay).
BER	Bit Error Rate.
B-ISDN	Broadband Integrated Services Digital Network.
CCITT	International Telegraph and Telephone Consultative Committee.
CD	Cepstral Distance.
CELP	Code Excited Linear Prediction.
C/R	Command/Response (of Frame Relay).
CRC	Cyclic Redundancy Check.
DE	Discard Eligibility (of Frame Relay).
DLCI	Data Link Connection Identifier (of Frame Relay).
DSPS	Double Sided Periodic Substitution.
DTX	Discontinuous Transmission.
DVCC	Digital Verification Color Code.
EA	Extended Address (of Frame Relay).
ETSI	European Telecommunications Standards Institute.

FACCH	Fast Associated Control Channel.
FCS	Frame Check Sequence (of Frame Relay).
FECN	Forward Error Correction Notification (of Frame Relay).
FER	Frame Error Rate.
FIR	Finite Impulse Response.
FLAT	Fast Lattice Algorithm (of EIA/TIA IS.54 VSELP).
FM	Frequency Modulation.
GSM	Groupe Spécial Mobile.
ICELP	Integer Code Excited Linear Prediction.
IS-54	Interim Standard - 54 for U.S. Digital Cellular Radio.
ISDN	Integrated Services Digital Network.
ITU	International Telecommunication Union.
LAR	Log Area Ratio.
LD-CELP	Low-Delay CELP (ITU G.728)
LP, LPC	Linear Prediction / Linear Predictor, Linear Predictive Coding.
LR	Likelihood Ratio.
LSF, LSP	Line Spectral Frequency, Line Spectrum Pair.
LTP	Long Term Predictor.
LSB	Least Significant Bit.
LWR	Levinson-Wiggins-Robinson Recursion.
MMSE	Minimum Mean Squared Error.
MOS	Mean Opinion Score.
MNRU	Modulated Noise Reference Unit.
MPE	Multi-pulse Excitation.
MSB	Most Significant Bit.
MS, MSE	Mean Squared, Mean Squared Error.
MSVQ, VQ	Multi-Stage Vector Quantization, Vector Quantization / Quantizer.

OSI	Open Systems Interconnection.
PCM	Pulse Coded Modulation (ITU G.711).
PL	Playback Length.
PR	Priority Bit.
PSTN	Public Services Telephone Network.
RPE-LTP	Regular Pulse Excited Long Term Prediction (ETSI GSM 06.10).
SACCH	Slow Associated Control Channel.
SD	Log-Spectral Distance.
SI, SID	Silence Indication Bit, Silence Descriptor.
SN	Sequence Number.
SNR	Signal to Noise Ratio.
SVD	Singular Value Decomposition.
TDMA	Time Division Multiple Access.
TIMIT	Texas Instruments / Massachusetts Institute of Technology Speech Database
TSM	Time Scale Modification.
VAD	Voice Activity Detection.
VLP	Vector Linear Prediction / Vector Linear Predictor.
VSELP	Vector Sum Excited Linear Prediction (EIA/TIA IS-54).
WMSE	Weighted Mean Squared Error.
ZIR	Zero Input Response.
ZSR	Zero State Response.

List of Symbols

$A, \bar{A}$	VLP coefficient supervector and supervector estimate.
A_j	VLP coefficient matrix employed in VLP.
$A_{1s,i}^{(g)}, A_{2s,i}^{(g)}$	VLP coefficient matrices for 1-sided and 2-sided prediction of stochastic gain vector.
$a, \hat{a}_n$	LPC coefficient column vector and the n^{th} LPC coefficient vector estimate.
$a', \hat{a}'_n$	Augmented LPC coefficient column vector and the n^{th} augmented vector estimate.
$A_n(z), A(z)$	Short-term LPC inverse filter.
$a_{n,k}$	The k^{th} LP coefficient of the n^{th} subframe.
$a_{1s,i}^{(f)}, a_{2s,i}^{(f)}$	Scalar LP coefficients for 1-sided and 2-sided prediction of frame gain.
α, α_p	Residual energy of $d_{n,k}$.
B_a	Actual buildout delay.
B_T	Target buildout delay.
$B_n(z), B(z)$	Long term predictor or adaptive codebook.
$b^{(m)}$	LTP or adaptive codebook coefficient vector.
$b_{n,k}, b$	The k^{th} LTP coefficient of the n^{th} subframe (of CELP) and LTP gain (of RPE-LTP).

β	Adaptive codebook gain (of VSELP).
$c_{k,j}$	The j^{th} basis vector of the k^{th} codebook (of VSELP).
c_i	The i^{th} cepstral coefficient.
$\mathbf{D}^{(m)}$	Residual excitation vector matrix.
D	Single bit delay employed in generation of convolution code.
$\hat{\mathbf{d}}^{(m)}, \hat{\mathbf{d}}^{(l)}$	Synthetic residual excitation vector contribution due to stochastic codeword l and LTP memory indexed by pitch lag m .
$d_{n,k}, \hat{d}_{n,k}$	Open loop residual after filtering $s_{n,k}$ by $A(z)$. The second term is the synthesized residual at the speech decoder.
$\hat{d}'^{(m)}(n), \hat{d}'_p^{(m)}(n)$	Residual excitation and partial residual excitation after filtering by H .
$d_{\text{LR}}(\mathbf{a}', \mathbf{a}'_p)$	Likelihood ratio between two LPC vectors.
$\bar{d}_{\text{LR}}$	Average likelihood ratio.
E_n	Total speech energy of the n^{th} subframe.
$\hat{e}_{n,k}^{(l)}$	Gain-shape stochastic excitation.
$\epsilon^{(l,m)}, \epsilon^{(l)}, \epsilon^{(m)}$	Weighted squared errors due to stochastic codeword l and/or pitch lag m .
ϵ_N	total power in the grid N excitation signal (of RPE-LTP).
$\mathbf{f}_{1,i}, \mathbf{f}_{2,i}$	filtered orthogonalized basis vectors (of VSELP).
$f_n, \hat{f}_n$	Frame gain and frame gain estimate (of ICELP).
f_s	Sampling frequency.
$G_n^{(l)}$	Stochastic gain scalar for codeword l of subframe n .
GS-P0-P1	Codevector of the transformed stochastic and adaptive codebook gains (of VSELP).
GS	Residual energy offset (of VSELP).
$g_m(n), g_{m \oplus L}(n)$	Polyphase filters for sample interpolation and $M: L$ rate conversion.

$\mathbf{g}_n, \hat{\mathbf{g}}_n$	Stochastic gain vector and associated estimate (of ICELP).
γ	Bandwidth expansion factor.
γ_1, γ_2	Stochastic codebook gains (of VSELP).
$\mathbf{H}, \mathbf{h}$	Lower triangular truncated impulse response matrix of $W_n(z)/A_n(z)$ and $\mathbf{h}$ is the first column of $\mathbf{H}$.
$h_a, \bar{h}_a$	Actual holding delay and mean actual holding delay.
$h_T, \bar{h}_T$	Target holding delay and mean target holding delay.
K_c	Length of subframe.
K_r	Length of truncated impulse response of $\mathbf{h}$.
$\mathbf{k}_n$	Reflection coefficient vector.
$k_{i,n}$	The i^{th} Reflection coefficient.
L	Interpolation factor.
l	Stochastic codebook index.
l_c	Total number of codevectors in a stochastic codebook.
$\Lambda_{n-1, n-2}$	Transformed matrix with $\mathbf{a}_{n-1}$ and $\mathbf{a}_{n-2}$ employed in likelihood prediction.
$\mathbf{M}_{1s,i}, \mathbf{M}_{2s,i}$	VLP coefficient matrices for 1-sided and 2-sided prediction of LPC coefficient vector.
M	Optimal pitch lag; decimation factor.
m	Candidate pitch lag.
$\mathbf{m}$	Likelihood LP vector.
N	Polyphase filter order or VLP order.
$\theta_{i,j}$	Sign multiplier of VSELP basis vector.
$P0, P1$	fractional values denoting respective energy contribution of adaptive codebook and stochastic codebook 1 towards R (in VSELP).
p	Order of the short-term LPC filter.

$p_i, p_{r,i}$	The i^{th} original and reconstructed speech parameter.
$p_k(n)$	Polyphase filter for decimation.
q	Order of LTP is $(m + q)$; queueing delay.
$\bar{q}, \bar{q}_{\max}$	Mean queueing delay and mean maximum queueing delay.
$\mathbf{R}, \tilde{\mathbf{R}}$	Autocorrelation matrix of $\mathbf{h}$ or autocorrelation supermatrix and supermatrix estimate.
$\mathbf{R}_p$	Autocorrelation matrix of a p^{th} order short-term LPC filter.
$\mathbf{R}_{ij}, \tilde{\mathbf{R}}_{ij}$	Autocorrelation coefficient matrix and estimate.
$R_q(0)$	Average quantized energy of a speech frame (in VSELP).
R	Total effective energy (of VSELP).
RS	Total energy of residual excitation $\hat{d}(n)$ (in VSELP).
$\mathbf{r}, \mathbf{r}^{(l)}$	Autocorrelation coefficient vectors of $\mathbf{h}$ and $\mathbf{v}^{(l)}$.
$\mathbf{r}_p$	Autocorrelation vector of a p^{th} order LPC filter.
$r(n)$	The n^{th} autocorrelation coefficient.
$s_{n,k}, \hat{s}_{n,k}$	Original and synthetic k^{th} speech sample in the n^{th} subframe.
σ	Noise weighting factor (in VLP), gain factor of LPC filter.
$\mathbf{u}$	ZIR vector of $W_n(z)/A_n(z)$.
$\mathbf{u}_{n-1,n-2}, \mathbf{U}_{n-1,n-2}$	Transformed “autocorrelation” vector and matrix employed in likelihood ratio prediction.
$\mathbf{V}, \tilde{\mathbf{V}}$	Autocorrelation supervector and estimate.
$\mathbf{v}_n^{(l)}, \mathbf{v}^{(l)}$	Stochastic codevector for codeword l of subframe n .
$v_{1I}(n), v_{2H}(n)$	Samples of stochastic codewords I and H (of VSELP).
$W_n(z)$	Perceptual weighting filter response.
$\mathbf{y}$	Weighted input speech vector i.e. speech filtered by $W_n(z)$.
$\hat{\mathbf{y}}^{(l)}, \hat{\mathbf{y}}^{(m)}$	ZSR of combined short-term LP and LTP filters and ZIR of LTP being fed into $W_n(z)/A_n(z)$.

ω	Angular frequency.
ω_n	LSP vector.
$\omega_{i,n}$	The i^{th} LSF.
$\omega_c, \omega_p, \omega_s$	Signal bandwidth, passband and stopband frequencies of polyphase filters.
$x_{\max}, x_M$	Block maxima and RPE samples employed in RPE-LTP.
$\otimes$	Convolution operator.
$\oplus$	modulo operator.

Chapter 1

Introduction

1.1 Background: Digital Speech Coding

Viewed as the most popular way of modern day communication, voice telephony continues to dominate all other services as the global telecommunication network evolves towards an Integrated Services Digital Network (ISDN) [42]. The trend towards digital telephony implies analog voice must be encoded in a digital format prior to transmission. Existing ITU (formerly CCITT) speech coding standards include G.711 64 kb/s μ /A-law Pulse Coded Modulation (PCM), G.726 32 kb/s Adaptive Differential Pulse Coded Modulation (ADPCM) and G.728 16 kb/s Low-Delay Code Excited Linear Prediction (LD-CELP).

Although the above bit rates suffice for most applications, recent development in mobile and packet communication networks shows there is a need for coding speech at lower bit rates. In the digital mobile radio environment, only a limited number of frequency bands has been allocated by the regulatory body. An efficient use of bandwidth can be achieved by multiplexing several low bit rate voice channels on a carrier frequency. The U.S. IS-54 digital cellular standard therefore specifies an 8 kb/s speech coder [14] while the European Groupe Spécial Mobile (GSM) system defines a similar coder at 13.2 kb/s [16]. For voice applications in the frame relay network [5], the transmission rate can vary greatly depending upon the network traffic condition. Statistical multiplexing employed in such packet switched networks means that no fixed bandwidth is allocated to any individual service. Low bit rate speech coding is beneficial because reduced packet

size increases queue service rate during congestion periods. At the same time a decreased bandwidth enables better utilization of other network resources.

In consideration of low bit rate and high speech quality requirements, the class of analysis-by-synthesis delayed decision coding schemes which include the Code Excited Linear Prediction (CELP) and the Multi-Pulse Excited Linear Predictive Coding (MPE-LPC) algorithms has been shown to work well at rates between 4.8 kb/s and 16 kb/s [33]. Analysis-by-synthesis delayed decision coding is a hybrid parametric and waveform coding scheme whereby a vocal tract model is employed to encode speech in a way that the decoded signal closely resembles the original speech waveform. Under this coding scheme, an excitation (or innovation) sequence is selected according to a weighted mean squared error distortion criteria and used to excite a set of cascaded time-varying linear filters. The low bit rate is achieved since only speech parameters representing the excitation sequence and the adaptive filters need to be transmitted.

As a variant of the CELP algorithm, the IS-54 Vector Sum Excited Linear Prediction (VSELP) coder uses a highly structured codebook to represent the excitation signal. On the other hand, the GSM Regular Pulse Excited Long Term Prediction (RPE-LTP) coder which is a special case of MPE-LPC employs a equal-spaced pulse train as its excitation source.

Besides having high speech quality at a low bit rate, a speech coder should be able to cope with a wide range of conditions. In the mobile environment, fading can cause bit as well as frame errors. Although Gray coding, bit interleaving and channel coding techniques can be employed to combat random bit errors, they are less effective for frame errors which can cause total corruption of the received parameters under deep fade conditions. Use of incorrect speech parameters can lead to substantial degradation in voice quality. A similar situation can arise in a congested packet switched network where voice packets can be dropped because of excessive network delay. Output speech quality can be impacted heavily if the lost packets are skipped during speech decoding [27].

1.2 Thesis Objective

The main objective of this thesis is to develop better voice frame reconstruction methods for use with CELP speech coders. Frame reconstruction methods are to be

employed when coded speech is transmitted over communication channels with finite frame loss rates. The methods should be designed to have low complexity, small processing delay and easily adapted for use with various CELP coding schemes. Most of all, the methods should enhance robustness of the CELP coders in the frame relay and digital cellular radio applications where frame loss is frequent.

In order to accomplish the above, this study requires the following:

- Understanding the basic frame relay protocol and hence adaptation for packet voice service;
- Understanding the basic analysis-by-synthesis coding framework and hence operation of the CELP speech coder;
- Design of a packet voice receiver for the frame relay network based on a CELP coder variant called Integer-CELP;
- Investigation of voice activity detection and time-scale modification methods as part of the packet voice receiver design;
- Understanding the speech coder requirement and error control mechanism currently being used in the IS-54 and GSM standards;
- Investigation of linear and non-linear estimation techniques suitable for reconstructing lost frame parameters; and
- Evaluation of speech quality against frame loss errors using various objective and subjective measurements.

1.3 Thesis Organization

The remainder of this thesis is organized into five main chapters and a conclusion. In Chapter 2, the frame relay network is introduced and problems associated with packet voice transmission are highlighted. The analysis-by-synthesis method of CELP speech coding is examined and Integer-CELP is then described as a suitable coder for sending speech over the frame relay network. To cope with variable network delay, an adaptive buildout delay strategy is described in association with voice activity detection and time-

scale modification of speech.

A second important application concerning speech coding over the digital cellular radio network is presented in Chapter 3. The basic architectures of the IS-54 and GSM systems are discussed. As robust speech coding is essential to enable reliable speech transmission over the mobile radio channel, the IS-54 VSELP and GSM RPE-LTP coding schemes are examined. To further protect the speech coders, error control mechanisms such as bit interleaving, cyclic redundancy check, channel coding and frame repetition are described.

Chapter 4 begins with a review of previous research on voice packet reconstruction. Waveform substitution, pitch detection, and embedded coding (with priority discarding) are described for PCM, ADPCM, Adaptive Predictive Coding (APC) and CELP speech coders. Utilizing the theory of Vector Linear Prediction (VLP), parametric extrapolation and interpolation schemes are then introduced for a number of CELP parameters including the short-term LPC filter representation, frame gain and stochastic excitation gain. Autocorrelation and covariance solutions to the VLP problem are explained. An alternative VLP formulation based on minimizing the likelihood ratio is then presented to estimate the lost LPC parameters. It is shown that the VLP method is less suitable for the adaptive codebook parameters, namely the pitch lag and the adaptive codebook coefficients. Instead a reconstruction method combining restricted pitch search at the encoder and median filtering at the decoder is used to estimate the lost pitch lags. Additionally, a conditional repeat strategy is described for the purpose of reconstructing the adaptive codebook coefficients. An adaptive energy control mechanism is introduced at the end of Chapter 4 to correct energy mismatch due to reconstruction errors.

In Chapter 5, standard methods of evaluating speech quality are reviewed. Objective and subjective results based on parametric extrapolation and interpolation are then presented at various frame error rates. Chapter 6 concludes the findings of this research and suggests possible areas to further improve the speech frame reconstruction method.

Chapter 2

Applications of Speech Coding (Part I): Frame Relay Networks

2.1 Introduction

In a circuit switched network like the Public Services Telephone Network (PSTN), a fixed bandwidth is always allocated for a communication channel. Although circuit switching is ideal for fixed rate voice communication, it is less suitable for variable rate services. For applications such as data communication, the average traffic requirement is usually less than the fixed channel capacity and hence circuit switched network can become under-utilized on a lot of occasions. Packet switching can be used instead to handle the variable rate traffic.

Frame relay is an example of such a packet switched network deployed in Broadband ISDN (B-ISDN) [5]. Without relying on any flow control mechanism, the frame relay network is a “bandwidth-on-demand” system which transfers information in variable length “frames”. If the sending device has a large burst of data to transmit, the frame relay interface then assigns adequate data bandwidth to enable the information be sent. If the sender is idling, the bandwidth can be allocated for other sending services within the network.

Embedded in the frame relay design is the concept of “fast packet”. In a conventional data communication network, continuous presence of a connection must be checked by exchanging of test information between the two end points as well as the intermediate

network nodes at all times. This overhead is eliminated by frame relay. Instead the sender only needs to place a frame marked with a destination address into the network, the data frame can be routed to the destination quickly through a high speed and low error-rate digital transmission media such as optical fiber.

In this chapter, speech coding over frame relay is examined. First of all, a brief overview of the frame relay protocol is given. The CELP algorithm which is used in many low-bit rate speech coding applications is then explained. Key elements of a packet voice receiver suitable for use in frame relay are discussed based on suitable modifications to the CELP algorithm.

2.2 Frame Relay Network Protocol

Figure 2.1 depicts the frame relay protocol with respect to the seven Open Systems Interconnection (OSI) layers. The frame relay User Network Interface (UNI) comprises of functions in the physical and data link layer. The physical layer is primarily responsible for the physical link between an attached station and the link which attaches that station to the network switching node. Once the physical link is established, the data link layer can transfer the data across the physical link in a sequence of frames. Specifically, the data link layer employs the CCITT Link Access Procedure-Frame Relay (LAP-F) or Link Access Procedure-Extended (LAP-E) to provide the following layer 2 core functions [5]: (1) frame delimiting, (2) alignment and transparency, (3) frame multiplexing/demultiplexing using the address field, (4) frame inspection to ensure the frame length is correct, (5) detection of transmission errors via Frame Check Sequence (FCS) and (6) congestion control via the Forward/Backward Explicit Congestion Notification (FECN/BECN) bits.

The frame relay frame structure is shown in Fig. 2.2. The fields in each frame are formatted and defined according to below [13]:

- **Flags:** use the bit pattern “01111110” as the frame delimiters. To ensure no bits in the user information field have the same pattern as the delimiting flags, “0” bit insertion takes place at the frame relay access device after any sequence with five consecutive “1”s. These extra “0” bits are subsequently removed at the destination.
- **DLCI:** The Data Link Connection Identifier is the address field in the frame relay

Figure 2.1 - Frame Relay Protocol Model

Figure 2.2 - Frame Relay Frame Structure

frame. The four octet version of DLCI is shown in Fig. 2.2. Normally, only the 2-octet format with 10 DLCI and 2 EA bits is employed and the 3-octet and 4-octet fields are eliminated. The Extended Address (EA) bits enable the frame-relay address field to extend from the default 1024 (10 bits/2-octet) addresses to 65,536 (16 bits/ 3-octet) and 8,388,608 (23-bit/4-octet) addresses.

A number of the DLCIs is reserved for exclusive use by the network. This includes DLCI 0 for channel call-control signalling, DLCI 1023 for Local Management Interface (LMI) communication and DLCIs 1-15 and 1008-1022 for future expansion. The remaining 992 DLCIs (numbered 16 to 1007) are for general use.

- **FECN and BECN:** The Forward and Backward Explicit Congestion Notification bits may be set by a congested network to notify the frame relay access device. The state table below clarifies the exact FECN and BECN bit settings. For example, if traffic congestion occurs only in the A→B direction but not the opposite direction, then a frame travelling from A→B will have the FECN bit set and BECN bit cleared. A packet in the B→A return path will have the FECN bit cleared and BECN bit set. These two bits are particularly useful in monitoring congestion during a (two-way) voice conversation.

Traffic Condition	Sending Direction	FECN	BECN
No congestion in either direction	A→B	0	0
	B→A	0	0
Congestion in A→B only	A→B	1	0
	B→A	0	1
Congestion in B→A only	A→B	0	1
	B→A	1	0
Congestion in both directions	A→B	1	1
	B→A	1	1

Table 2.1 - State Table for FECN and BECN Bits

- **DE:** The Discard Eligibility bit can be used in conjunction with a voice packet prioritization protocol (- see Sections 4.1.4 and 4.1.5). If the DE bit is set for a particular frame, the frame relay network can choose to discard this frame during

congestion. Frames which have DE bit cleared should not be discarded.

- **EA:** Extended Address bits are the LSBs in the second to fifth byte of the frame. A 10-bit DLCI address is defined with the EA bit (of the second byte) set to 0 and EA bit (of the third byte) set to 1. See also **DLCI**.
- **C/R:** The Command/Response indication bit is not used by the frame-relay data link layer protocol and is carried transparently by the frame relay network.
- **FCS:** Frame Check Sequence is used to verify a frame has not been corrupted during transmission. It consists of a two-byte field as a result of applying the CCITT Cyclic Redundancy Check (CRC) polynomial to the frame from the first bit of the DLCI address up to the last bit of the user data field. At the destination, the FCS is recalculated and if it does not match with the received FCS, the frame is discarded.
- **User Information:** Either user or system control information can be placed in this variable length field. Depending upon the efficiency of the speech coding algorithm as well as trade-offs with the frame relay overhead, several sets of speech parameters may be packed into one frame relay frame.

2.3 Speech Coding in the Frame Relay Network

At the frame relay UNI, input speech samples are encoded into a set of compact parameters and sent over the frame relay network. As the voice frames traverse across the network, they are subjected to the following effects:

- **Delay:** The total network delay is comprised of (1) speech coding delay, (2) transmission time it takes for the frame relay network device to place the frames on the links, (3) signal propagation delay, (4) switching time and (5) queueing delay q [13]. Whereas the first four delay components are fixed in values, the queueing delay can cause significant variation in the total network delay. Since decoded speech must be played out in a synchronous manner, it is necessary to insert some kind of buffering (called the actual holding delay h_a) at the receiver to compensate for this unknown value of q . The actual buildout delay B_a is given by: $B_a = q + h_a$. Frame relay frames which have buildout delay exceeding the preset threshold (called the target buildout delay B_T) are dropped.

- **Transmission Error:** Although bit errors are less frequent in the frame relay network (because of the reliable transmission media), they can cause the packet voice receiver to fail to decode a header sequence and results in an out-of-frame condition.
- **Traffic Congestion:** Speech frames may be lost due to buffer overflow or they may be dropped in the network nodes by the congestion control mechanism. Packet discarding can also occur if the number of frames is generated in excess of transmission capacity at the source.
- **Echoes:** Echo cancellation may be necessary if the total round-trip network delay exceeds beyond the acceptable limit for a voice conversation.

In this discussion, the issues on variable network delay and network congestion are addressed. Since both of these network effects have direct impacts on the speech coder, the CELP coding algorithm is next explained. A delay equalization strategy combining voice activity detection and time-scale modification methods is then described. Several speech frame reconstruction procedures to cope with frame losses during congestion periods are introduced towards the end of this chapter.

2.3.1 Code Excited Linear Prediction (CELP)

This section provides a quick overview of the CELP algorithm. Since the subject is treated here in general terms, readers may want to refer to [34][35] for a complete discussion.

The basic structure of the analysis-by-synthesis CELP coder is illustrated in Fig. 2.3. As indicated in the diagram, there is a synthesis as well as an analysis block to the structure. The synthesis block (which also acts as the speech decoder) is based on the well-known linear time-varying speech production model as shown in Fig. 2.4. Synthetic speech $\hat{s}_{n,k}$ is hence generated by a Gaussian noise source exciting two all-pole adaptive linear filters. The short-term Linear Predictive (LP) filter $1/A_n(z)$ which models the near-sample correlation or speech spectral envelope is defined as:

$$\frac{1}{A_n(z)} = \frac{1}{1 + \sum_{k=1}^p a_{n,k} z^{-k}} \quad (2.1)$$

Figure 2.3 - Analysis-by-Synthesis CELP Speech Coder Structure

Figure 2.4 - Analysis-Synthesis Filtering of Speech

The denominator $A_n(z)$ is a p^{th} order LP inverse filter with time-varying coefficients $a_{n,k}$. Typical values of p are between 10 and 16 assuming an 8 kHz sampling rate. In general, the set of filter coefficients is designed to minimize the prediction error obtained by filtering the original speech sample $s_{n,k}$ through $A_n(z)$. Solution to this minimization process and hence the set of filter coefficients $\{a_{n,k}\}$ is discussed as a special case in Section 4.2.1.

The Long-Term Predictive (LTP) filter $1/B_n(z)$ which models the far-sample correlation or spectral fine structure of the speech signal is represented by:

$$\frac{1}{B_n(z)} = \frac{1}{1 - \sum_{k=-q}^q b_{n,k} z^{-(m+k)}} \quad (2.2)$$

where m is a particular lag with typical values between 16 and 150 sampling periods[†] and $\{b_{n,k}\}$ are the time-varying LTP coefficients. Although the order of inverse filter $B_n(z)$ is large, only $(2q+1)$ coefficients have non-zero values and q is usually chosen to be either 0 or 1. The lag and filter coefficients can be determined either in an open loop configuration by minimizing the prediction error $e_{n,k}$ obtained by filtering $d_{n,k}$ through $B_n(z)$ (Fig. 2.4(a)) or closed loop using the analysis-by-synthesis CELP structure in Fig. 2.3. The advantage of determining m and $b_{n,k}$ in a closed-loop configuration is that the LTP parameters are contributing to the minimization of the Weighted Mean Squared Error (WMSE) $\epsilon_n^{(l)}$. As a result, performance of the closed-loop LTP filter is significantly better than its open loop counterpart. The closed-loop minimization process is described in Section 2.3.1.1.

In the analysis stage, the CELP coder “windows” input speech samples into a series of analysis frames which are further partitioned into subframes of length K_c . Adopting the notation in [35], the k^{th} original and synthetic speech samples in the n^{th} subframe are denoted by $s_{n,k}$ and $\hat{s}_{n,k}$ respectively. The subscript n may be dropped where no confusion may result. The difference between $s_{n,k}$ and $\hat{s}_{n,k}$ is perceptually weighted by a filter $W_n(z)$

[†] Fractional rate long-term LP filter also exists whereby the lag period takes on non-integral values through sample rate conversion. A single-tap fractional rate LP filter is reported to yield improved prediction gain for high-pitch speakers over conventional multi-tap integral rate long-term LP filter. See [32] for details.

defined as:

$$W_n(z) = \frac{A_{n,p}(z/\gamma_1)}{A_{n,p}(z/\gamma_2)} \quad (2.3)$$

where $A_{n,p}(z)$ denotes the unquantized short-term p^{th} order LP inverse filter defined in (2.1). The bandwidth expansion factors γ_1 and γ_2 (between 0 and 1) determine the amount of de-emphasis applied to the formant regions in the error spectrum. In this way, a more uniform signal to noise ratio could be obtained across all frequency regions.

For a particular subframe, the WMSE $\epsilon_n^{(l)}$ is used as a distortion measure to select a codeword l (from a codebook) representing a stochastic codevector $\mathbf{v}_n^{(l)}$. This "shape" codevector is scaled by a gain scalar $G_n^{(l)}$ and then fed into a cascade of long-term and short-term LP filters to synthesize the speech. Referring to Fig. 2.3, the k^{th} synthetic speech sample, stochastic excitation codevector, gain-scaled codevector and residual excitation elements which are generated as a result of the codeword l selection are denoted by $\hat{s}_{n,k}^{(l)}$, $\mathbf{v}_{n,k}^{(l)}$, $\hat{\mathbf{e}}_{n,k}^{(l)}$ and $\hat{\mathbf{d}}_{n,k}^{(l)}$ respectively.

2.3.1.1 Complexity Reduction and Search Strategies

In the original CELP speech coder [51], the optimum codeword l which yields the minimum WMSE is obtained by an exhaustive search of the stochastic codebook. Thus for each codeword l , $s_{n,k}$ must be compared with the synthetic speech $\hat{s}_{n,k}^{(l)}$ generated by passing the gain-scaled excitation $\hat{\mathbf{e}}_{n,k}^{(l)}$ through the LP filters and the difference signal is then weighted by $W_n(z)$ prior to WMSE computation. This leads to immense computational cost and prohibits direct implementation of the CELP algorithm. Alternative solutions with reduced complexity must be found.

The first step of complexity reduction is to rearrange the CELP coder structure according to Fig. 2.5 [35]. The weighting filter $W_n(z)$ is moved before the summing node and the synthetic speech is decomposed into a Zero-State Response (ZSR) signal and a Zero-Input Response (ZIR) signal. ZSR refers to a LP filter response with all zero start-up filter memory whereas ZIR is the response caused by memory hangover. Through this simplification, the WMSE (sans a minor scale factor) can be written as:

$$\epsilon^{(l,m)} = |\mathbf{y} - \mathbf{u} - \hat{\mathbf{y}}^{(l)} - \hat{\mathbf{y}}^{(m)}|^2 \quad (2.4)$$

Figure 2.5 - Rearranged CELP Coder Structure

(1) Perceptual Weighting Filter $W_n(z)$ is moved prior to summation and (2) the ZIR and ZSR are separated for each filter.

where y is the weighted speech input, u is the ZIR vector of the combined $W_n(z)/A_n(z)$ filter, $\hat{y}^{(l)}$ is the ZSR vector of the cascaded synthesis filters and $\hat{y}^{(m)}$ is the ZIR vector of the long-term LP filter memory. If the truncated impulse response of $W_n(z)/A_n(z)$ is denoted by the column vector h and H is a lower triangular Toeplitz matrix with h in its first column, then (2.4) can be expressed in terms of the following:

$$\epsilon^{(l,m)} = |y - u - H\hat{d}^{(m)} + H\hat{d}^{(l)}|^2 \quad (2.5)$$

Closed loop joint optimization of stochastic codebook index l , lag m and the LTP coefficients $b_{n,k}$ is exceedingly complex and a sequential search may be used. First, m and $b_{n,k}^{(m)}$ which result in minimum WMSE are obtained by letting the stochastic codebook gain $G_n^{(l)}$ be zero such that $\hat{d}^{(l)} = 0$. After the optimal values M and $b_{n,k}^{(M)}$ have been found, the stochastic codeword l and associated gain $G_n^{(l)}$ can be determined.

Assuming $m \geq K_c + q$, the objective function of (2.5) can be further simplified by expressing the residual excitation $\hat{d}^{(m)}$ as:

$$\hat{d}^{(m)} = G^{(l)} v^{(l)} + \sum_{j=-q}^q b_j \hat{d}_{-(m+j)} \quad (2.6)$$

The above equation can be written compactly as:

$$\hat{d}^{(m)} = G^{(l)} v^{(l)} + D^{(m)} b^{(m)} \quad (2.7)$$

where

$$b^{(m)} = [b_{-q}^{(m)} \dots b_0^{(m)} \dots b_q^{(m)}]^T \quad (2.8)$$

$$\hat{d}_{-(m+j)} = [\hat{d}_{-(m+j)} \hat{d}_{1-(m+j)} \dots \hat{d}_{K_c-1-(m+j)}]^T \quad (2.9)$$

and $D^{(m)}$ is a $K_c \times 2q+1$ matrix given by:

$$D^{(m)} = [\hat{d}_{q-m} \hat{d}_{q-1-m} \dots \hat{d}_{-m-q}] \quad (2.10)$$

By letting $G_n^{(l)} = 0$, $\epsilon^{(l,m)}$ can be expressed solely in terms of m such that:

$$\begin{aligned}\epsilon^{(m)} &= \left| (y - u) - \mathbf{H}\mathbf{D}^{(m)} \mathbf{b}^{(m)} \right|^2 \\ &= \mathbf{y}_1^T \mathbf{y}_1 - 2\mathbf{y}_1^T \mathbf{H}\mathbf{D}^{(m)} \mathbf{b}^{(m)} + \mathbf{b}^{(m)T} \mathbf{D}^{(m)T} \mathbf{H}^T \mathbf{H} \mathbf{D}^{(m)} \mathbf{b}^{(m)}\end{aligned}\quad (2.11)$$

where

$$\mathbf{y}_1 = \mathbf{y} - \mathbf{u} \quad (2.12)$$

The optimal LTP coefficient vector (since $G_n^{(0)}=0$) can be obtained by minimizing $\epsilon^{(m)}$ with respect to $\mathbf{b}^{(m)}$ to yield:

$$\mathbf{b}^{(m)} = (\mathbf{D}^{(m)T} \mathbf{R} \mathbf{D}^{(m)})^{-1} \mathbf{D}^{(m)T} \mathbf{c}_1 \quad (2.13)$$

where $\mathbf{c}_1 = \mathbf{H}^T \mathbf{y}_1$ and $\mathbf{R} = \mathbf{H}^T \mathbf{H}$ is the autocorrelation matrix. The adaptive codebook coefficient vector, $\mathbf{b}^{(m)}$, can be substituted back in (2.11) to give:

$$\epsilon^{(m)} = \mathbf{y}_1^T \mathbf{y}_1 - \mathbf{c}_1^T \mathbf{D}^{(m)} (\mathbf{D}^{(m)T} \mathbf{R} \mathbf{D}^{(m)})^{-1} \mathbf{D}^{(m)T} \mathbf{c}_1 \quad (2.14)$$

To obtain the optimal $\epsilon^{(M)}$, (2.14) must be computed for every lag m . Even though the number of lag candidates is limited (to 128 values usually), the above expression is far too complex because of the computations involving the $2q+1$ dimensional vector $\mathbf{b}^{(m)}$. A sub-optimal solution is usually employed whereby a closed-loop search can be performed by assuming a single-coefficient LTP filter with $\mathbf{b}^{(m)} = b_0^{(m)}$. Through this assumption, (2.14) is simplified to:

$$\epsilon^{(m)} = \mathbf{y}_1^T \mathbf{y}_1 - \frac{(\mathbf{c}_1^T \mathbf{D}^{(m)})^2}{|\mathbf{H}\mathbf{D}^{(m)}|^2} \quad (2.15)$$

Efficient methods can be found in the evaluation of $\epsilon^{(m)}$ either by taking advantage of the memory recursion in $\mathbf{D}^{(m)}$ or imposing a Toeplitz structure [35] on the autocorrelation matrix $\mathbf{R}$ which has elements element $R_{i,i+j}$ given by:

$$R_{i,i+j} = r_j = \sum_{i=0}^{K_r-1} h_i h_{i+j} \quad (2.16)$$

where K_r is the length of truncated impulse response $\mathbf{h}$. (The truncation is applied to reduce computational complexity.) In both cases, the m^{th} value of the expression $|\mathbf{H}\mathbf{D}^{(m)}|^2 = \mathbf{D}^{(m)T}\mathbf{R}\mathbf{D}^{(m)}$ can be computed recursively from its $(m-1)^{\text{st}}$ result [35]. Since the lag M obtained from the above procedure is only a sub-optimal value, (2.13) cannot be used and the corresponding coefficient vector, $\mathbf{b}^{(M)}$, is found instead by minimizing the WMSE criteria defined in (2.11). In practice, $\mathbf{b}^{(M)}$ is chosen from a Vector Quantized (VQ) codebook containing a set of quantized coefficient vectors.

In deriving the residual excitation $\hat{\mathbf{d}}^{(m)}$, it is assumed the lag m is always greater than the quantity $K_c + q$ such that no high order and cross product terms of b_j can appear in $\epsilon^{(m)}$. To guarantee this condition is not violated by small values of m , (2.6) can be suitably modified as follows:

$$\hat{\mathbf{d}}^{(m)} = G^{(l)} \mathbf{v}^{(l)} + \sum_{j=-q}^q b_j \begin{cases} \hat{\mathbf{d}}_{-(m+j)}, & k-j < m \\ \hat{\mathbf{d}}_{-(2m+j)}, & m \leq k-j < 2m \\ \hat{\mathbf{d}}_{-(3m+j)}, & 2m \leq k-j < 3m \\ \dots & \dots \end{cases} \quad (2.17)$$

The above formulation defines the structure of an **Adaptive Codebook** (ACB). Thus in the closed-loop WMSE minimization, m becomes a codebook index pointing to a candidate codevector (or past residual excitation vector) in the LTP filter memory. The adaptation therefore refers to the memory shift which occurs in $\hat{\mathbf{d}}^{(m)}$ after each subframe synthesis. It can be seen (2.17) is identical to the LTP filter definition in (2.6) if $m \geq K_c + q$. Unless otherwise stated, the adaptive codebook definition is assumed in this discussion.

The remaining CELP parameters still to be resolved are the stochastic codebook index l and the associated codevector gain $G^{(l)}$. The task is made simpler by the fact the ACB parameters, M and $\mathbf{b}^{(M)}$, have already been found and can be eliminated from the WMSE ($\epsilon^{(l,m)}$). Noting that $\hat{\mathbf{d}}^{(l)} = G^{(l)} \mathbf{v}^{(l)}$, a simplified WMSE expression can be obtained by substituting the ACB parameters into (2.11) and combining with (2.5) to give:

$$\begin{aligned} \epsilon^{(l)} &= |(\mathbf{y} - \mathbf{u} - \mathbf{H}\mathbf{D}^{(m)} \mathbf{b}^{(m)}) - G^{(l)} \mathbf{H} \mathbf{v}^{(l)}|^2 \\ &= (\epsilon^{(M)} - 2G^{(l)} \mathbf{y}_2^T \mathbf{H} \mathbf{v}^{(l)} + G^{(l)2} \mathbf{v}^{(l)T} \mathbf{H}^T \mathbf{H} \mathbf{v}^{(l)}) \end{aligned} \quad (2.18)$$

where

$$\mathbf{y}_2 = \mathbf{y} - \mathbf{u} - \mathbf{H}\mathbf{D}^{(M)}\mathbf{b}^{(M)} \quad (2.19)$$

Optimal value of the stochastic gain can be found by minimizing $\epsilon^{(l)}$ with respect to $G^{(l)}$:

$$G^{(l)} = \frac{(\mathbf{c}_2^T \mathbf{v}^{(l)})^2}{\mathbf{v}^{(l)T} \mathbf{R} \mathbf{v}^{(l)}} \quad (2.20)$$

where $\mathbf{c}_2 = \mathbf{H}^T \mathbf{y}_2$ and by substituting $G^{(l)}$ back into (2.17), $\epsilon^{(l)}$ is reduced to:

$$\epsilon^{(l)} = \epsilon^{(M)} - \frac{(\mathbf{c}_2^T \mathbf{v}^{(l)})^2}{\mathbf{v}^{(l)T} \mathbf{R} \mathbf{v}^{(l)}} \quad (2.21)$$

As $\epsilon^{(M)}$ is fixed after computation of the adaptive codebook parameters, computational complexity of the above expression is largely dominated by the second term. In particular, it is noted the denominator of the second term has computational complexity proportional to K_c^2 . Several complexity reduction methods employed by the frame relay speech coder and the IS-54 VSELP coder are now explained.

Autocorrelation Approach - Similar to the computation of the lag value, the matrix $\mathbf{R}$ is first approximated by a more convenient Toeplitz form and then the denominator in (2.21) can be simplified to:

$$\mathbf{v}^{(l)T} \mathbf{R} \mathbf{v}^{(l)} = 2\mathbf{r}^T \mathbf{r}^{(l)} \quad (2.22)$$

where r_j is as defined in (2.16) and $r_j^{(l)}$ is essentially the autocorrelation coefficients of the stochastic codevector given by:

$$r_j^{(l)} = \begin{cases} \frac{1}{2} \sum_{i=0}^{K_c-1} (v_i^{(l)})^2 & j = 0 \\ \sum_{i=0}^{K_c-1-j} v_i^{(l)} v_{i+j}^{(l)} & j \neq 0 \end{cases} \quad (2.23)$$

The autocorrelation approach is employed by the frame relay coder.

Structured Codebooks - By imposing special structures in the codebook, significant reduction in codebook storage and search complexity can be obtained with little loss in performance degradation [35]. In a basis vector codebook design, the stochastic vector is expressed as the sum of a scaled set of basis vectors:

$$\mathbf{v}^{(l)} = \mathbf{C} \mathbf{b}^{(l)} \quad (2.24)$$

In the above, $\mathbf{b}^{(l)}$ identifies the codeword as well as the scale factors for individual basis vector and $\mathbf{C}$ designates a basis matrix given by:

$$\mathbf{C} = [\mathbf{c}_1 \ \mathbf{c}_2 \ \dots \ \mathbf{c}_N] \quad (2.25)$$

where $\mathbf{c}_1$ to $\mathbf{c}_N$ are the N basis column vectors of length K_c . If $\mathbf{b}^{(l)}$ is an L level binary code with corresponding ± 1 scale factors, then $N = \log_2 L$. Furthermore, if $\mathbf{b}^{(l)}$ is Gray coded such that $\mathbf{b}^{(l)}$ and $\mathbf{b}^{(l-1)}$ always differ by only a single bit, then (2.23) can be computed recursively. This particular approach is adopted by the IS-54 VSELP coder. Further explanation of the VSELP algorithm can be found in Section 3.3.1.

Multi-Stage Vector Quantized (MSVQ) codebook is another design approach adopted by the frame relay speech coder. As illustrated in Fig. 2.6, the stochastic excitation is formed by summing a series of codevectors selected from a number of smaller codebooks

$$\mathbf{v}^{(l_1, l_2, \dots, l_M)} = \sum_{i=1}^M \mathbf{v}_i^{(l_i)} \quad (2.26)$$

The index set $\{l_1, l_2, \dots, l_M\}$ is used to select M codevectors from the codebooks. A sub-optimal sequential search procedure (based on the autocorrelation approach) may be employed in which case (2.21) is first minimized by finding l_1 . Next, l_2 is found with l_1 fixed and the procedure continues until all M indices are obtained. The use of structured codebooks can be introduced to speed up the search. An M - L search procedure [35] which retains multiple candidates at each stage during the search has also been proposed for improved performance.

By making use of some of the above methods as well as other enhancements, a CELP speech coder suitable for frame relay was designed and forms the subject of the next

Figure 2.6 - MSVQ of CELP Stochastic Excitation

Figure 2.7 - ICELP Frame Buffering

section.

2.3.1.2 Integer-CELP: A Frame Relay Speech Coder

Parameter	Bits/subframe	Bits/frame
MSVQ short-term LP filter Indices	-	30
Frame Gain: F	-	4
ACB Lags: $M_1 - M_4$	7	28
Stochastic Excitation Gain: $G_1 - G_4$	4	16
ACB Coefficient Indices: $B_1 - B_4$	4	16
MSVQ Stochastic Codebook Indices: $V_1 - V_4$	16	64
Future Expansion Bits: E	-	2
TOTAL:	31	160

Table 2.2 - ICELP Frame Format

The reference coder used in this study is a fixed-point CELP variant called Integer-CELP (ICELP) capable of producing high quality speech at 8 kb/s. Specifically, ICELP is a closed-loop analysis-by-synthesis speech coder which consists of a short-term 10th order LP filter, a long-term LP filter (in the form of a 3rd order adaptive codebook), a perceptual weighting filter and two MSVQ stochastic codebooks. The short-term LP filter coefficients are represented by a 7-stage MSVQ. Details of MSVQ short-term filter design can be found in [35][36]. Benefits of the structured MSVQ codebook design include low encoding rate as well as robustness against transmission errors.

Key specifications of the ICELP coder are listed below:

1. The sampling frequency f_s is 8 kHz and the equivalent sampling period is 125 μ s.
2. Analysis frame length is 20 ms (160 samples). This is further divided up into four 5ms subframes each of 40 samples. A Hamming window is used to weight the input speech samples prior to LPC analysis. The analysis window is centered over the last subframe to provide a frame overlap of 100 previous samples and 60 current samples.

3. Since ICELP is a forward adaptive coder, an inherent delay of 220 samples (160 samples for the frame and 60 overlapped samples) is needed to process one speech frame. As shown in Fig. 2.7, the encoder buffers the n^{th} speech frame, processes this buffered frame in the $(n+1)^{\text{st}}$ frame and outputs the encoded parameter set in the $(n+2)^{\text{nd}}$ frame. At the receiver, the parameters can be decoded to synthesize the speech frames. Decoding delay is negligible compared to the encoding delay.
4. A 160-bit parameter set is generated every 20 ms as shown in Table 2.2. An ICELP-encoded frame consists of the following (in the order of transmission beginning with MSB): 2 future expansion bits (E), a 30-bit sequence of short-term LPC parameter MSVQ codebook indices, a 4-bit quantized frame gain (F) and four sets of subframe parameters. Each subframe then consists of a 7-bit lag value (M_i), a 4-bit ACB coefficient index (B_i), a 16-bit sequence of MSVQ stochastic excitation codewords (V_i) and a 4-bit quantized stochastic gain value (G_i).
5. VQ codebooks and quantization tables are used by ICELP for coding and decoding purposes. The F and G_i quantization tables have 16 entries. The M_i codebook has 128 integer entries with values from 21 to 147. The 3-dimensional ACB coefficient vector $\mathbf{b}^{(m)}$ is stored in a 16-level VQ codebook indexed by B_i . The stochastic excitation is formed by the sum of two 40-dimensional signed stochastic vectors selected from a 2-stage MSVQ codebook. Each stage has a 128-level (7-bit) codebook with a sign multiplier (1-bit). Together the two stages make up the 16-bit stochastic excitation codeword V_i . The short-term LP filter coefficients are represented by a series of MSVQ codebook indices of bit sizes: {6,4,4,4,4,4,4}. The seven indices are used to select the corresponding codevectors from one 64-level and six 16-level codebooks. During the decoding process, a 10-dimensional vector is selected from each codebook, summed together and transformed to obtain the required short-term filter LP coefficient vector $\mathbf{a}$. In Chapter 4, transformed parameters such as reflection coefficients and line spectrum pairs are considered.

As discussed earlier in Section 2.2, a frame relay frame contains 6 bytes of network signalling information including frame delimiters, DLCI, FECN/BECN, DE, EA, C/R and FCS. Except for the possible use of FECN/BECN and DE bits, the signalling information for voice frame transmission has to be carried in-band along with the speech parameters.

Figure 2.8 shows an “enhanced frame” structure suitable for use with the ICELP coding scheme. The standard frame relay frame is supplemented with the following fields:

- **SN:** the sequence number is a wrap-around n -bit counter which is used to determine the chronological order of received voice frames. A dropped packet will show up as a discontinuity in the sequence number.
- **PR:** the priority bit[†] is always set to 1.
- **SI:** the silence indicator bit is set if the current frame contains silence. In conjunction with the DE and PR bits, the SI bit is useful for congestion control in a frame relay voice protocol which allows silent frames to be dropped. Normally, the first silent frame (after a talkspurt) is sent with DE=0, SI=1, PR=0 and subsequent silent frames are sent with DE=1, SI=1, PR=0. For non-silent frames, the bits are set such that DE=0, PR=1, SI=0. The PR and SI bits may be embedded in the two E bits of the ICELP speech frame.

Additionally, a 2-bit Playback Length (PL) field is generated internally by the packet voice receiver to indicate to the ICELP decoder to adjust the speech playback rate. Five different playback lengths: {10ms, 15ms, 20ms, 25ms, 40ms} have been defined and they are selected for use depending upon whether the speech frame contains silence or non-silence. The main purpose of the PL bits is to provide a mechanism for network delay equalization. One such proposed delay equalization scheme called *adaptive buildout delay* [13] suitable for use in the frame relay network is herein explained.

At the beginning of a frame relay voice connection, a holding time which is either fixed or based on $(1/2 \times \text{round-trip delay measurement value})$ is inserted in between the arrival of the first packet and the time it is decoded. This holding time essentially provides an initial estimate of the buildout delay [41]. As the network traffic starts to vary, the buildout delay will change with time. Thus if a fixed holding time is used and is set too large initially, no packets are ever lost but the end-to-end delay will be excessive. On the other hand if the holding time is too small, too many packets will be dropped due to an

[†] In the 16 kb/s embedded CELP-MPE coder [37] design, the auxiliary 8 kb/s stream is sent in a separate frame and the PR bit is therefore used to identify the embedded stream from the auxiliary information. For the purpose of this present work, the ICELP coder is always assumed to be non-embedded.

Figure 2.8 - Frame Relay Speech Frame Structure

increase in buildout delay. Thus the buildout delay should be set such that an acceptable percentage of frames is lost.

The adaptive buildout delay mechanism attempts to adjust the holding time continuously until it reaches certain error bound from a target value called the target holding time over the long-term average. Denoting the mean target holding time, mean actual holding time and mean queueing delay respectively as $\bar{h}_T$, $\bar{h}_a$ and $\bar{q}$, it is possible to write:

$$\begin{aligned}\bar{h}_T &= B_T - \bar{q} \\ \bar{h}_a &= B_a - \bar{q}\end{aligned}\tag{2.27a,b}$$

where B_a and B_T are the actual and target buildout delays defined earlier. Since B_T is fixed, it has been shown in [13] $\bar{h}_T$ can be computed by replacing $\bar{q}$ with $\bar{q}_{\max}$ which is the mean maximum queueing delay. Value of $\bar{h}_a$ is approximated by the accumulated average of h_a and by comparing the value of $\bar{h}_a$ with $\bar{h}_T$, different time-scale adjustments can be made. In particular if $\bar{h}_a > \bar{h}_T$, the playback length of the speech frame is shortened and on the other hand if $\bar{h}_a < \bar{h}_T$, the display length is then lengthened. The target error bound $\epsilon < |\bar{h}_T - \bar{h}_a|$ can be used to judge whether the system has reached steady state after a certain number of adjustments.

2.3.2 Voice Activity Detection over Frame Relay

Voice Activity Detection (VAD) enables the frame relay network to prioritize speech frames accordingly. During congestion control, silent frames may be dropped. At the same time, identification of a silence period enables the above adaptive buildout delay mechanism to be used to equalize network delay.

For this particular study, a fixed-threshold energy detector (with hangover interval) was chosen since noise does not present a problem. For real-time system, better VAD like the one in GSM digital cellular [52] should be used. The energy detector computes a mean energy value at the end of a 20 ms analysis frame and compares this value with a fixed threshold. The speech frame is declared as silence (SI =1) if the signal power drops below the threshold level and this energy level has been maintained for a period of time greater than the preset hangover interval. If neither of the two conditions were met, the speech detector then declares the frame to be non-silence (SI =0). A 200 ms hangover interval

(equivalent to 10 speech frames) was chosen based on a suggestion in [25]. The silence threshold is computed as (average energy - 4 × standard deviation). The average energy and standard deviation were obtained from 25,000 speech frames of the TIMIT database.

2.3.3 Time-Scale Modification (TSM) of CELP Coded Speech

Five different conditions are specified by the PL bits for adaptive buildout delay adjustment. Since silent frames are perceptually less important, large playback adjustments (halving or doubling the nominal playback length) are possible. For non-silent frames, the playback adjustment is set to be $\pm 25\%$ of the nominal length. Upon receiving a speech frame, the packet voice receiver examines the PL and SI bits and determines suitable TSM according to the specification tabulated below:

PL Bits	SI Bits	Time-Scale Factor	Change in Playback Length	Comments
00	x	1:1	20ms (nominal)	No TSM is needed
01	0	4:5	20ms → 25ms	Time-scale expansion of non-silent speech frame
10	0	4:3	20ms → 15ms	Time-scale compression of non-silent speech frame
11	1	1:2	20ms → 40ms	Time-scale expansion of silent speech frame
10	1	2:1	40ms → 20ms	Time-scale compression of silent speech frame

Table 2.3 - TSM Specification of Received CELP Frame

Suitable time-scaling techniques must be derived to enable playback of speech frames at the above given rates. A review of literature shows that all time-scale modification methods which include linear filtering [1], short-time Fourier analysis [45] and sinusoidal representation [47] are based on availability of the original speech waveform. In order to make use of one of these methods, it would first require the speech decoder to synthesize the speech waveform before performing the TSM. It is also noted that the previous methods are aimed at reproducing high quality time-scaled modified speech on a continuous basis. The frame relay application differs by the fact that the adaptive buildout

delay mechanism time-scale modifies only a small number of non-contiguous frames during a voice connection. In this case, continuous TSM is not necessary because of the discrete and random nature of the frame adjustment. A simple approach based on the ICELP parametric framework is therefore considered.

Since the speech signal exhibits quasi-periodicity when it is voiced, the gain-scaled stochastic excitation vector $\hat{e}_n^{(i)}$ may be time-scale modified before feeding into the LTP and the short-term LPC filters. The excitation vector is either lengthened or shortened according to the appropriate time-scale factor. The lag M , ACB coefficient vector $b^{(M)}$ and the LPC coefficient vector a_n remain unmodified in order to preserve the spectral characteristics of the original speaker. The operation is illustrated in Fig. 2.9. In this study, two simple TSM methods which involve sampling rate conversion (via polyphase filtering) and sample replication and truncation were explored.

In the sampling rate conversion approach, the decoded excitation vector $\hat{e}_n^{(i)}$ is fed into one of several polyphase filters selected according to the received values of the PL and SI bits. The rate-converted samples are then used directly as time-scaled samples in the speech frame. For example in the time-scale expansion $20\text{ms} \rightarrow 25\text{ms}$, the 160 $\hat{e}_{n,k}^{(i)}$ samples at $125 \mu\text{s}$ spacing (or 8 kHz sampling frequency) are converted to 200 $\hat{e}'_{n,k}^{(i)}$ samples at $100 \mu\text{s}$ spacing (or 10 kHz sampling frequency). The rate converted $\hat{e}'_{n,k}^{(i)}$ are then used as the “effective” time-scaled excitation samples[†].

To generate the required number of time-scaled samples, the TSM operations defined in Table 2.3 are mapped to the sampling rate conversion schemes. Sample decimation and interpolation are employed respectively to obtain the 2:1 and 1:2 TSMs whereas fractional rate converters are used to obtain the 4:5 and 4:3 TSMs. The three operations are depicted in Fig. 2.10 (a)-(c). Appendix A provides a brief discussion on sampling rate conversion theory and practical implementation via polyphase filtering

An alternative TSM method can be derived by simple replication or truncation of the stochastic excitation vector. For example in the 4:5 TSM, 1/4 of the original stochastic vector can be repeated and appended to the original vector to provide the time-scaled

[†] It should be cautioned sampling rate conversion and time-scale modification are two different physical operations. On this particular occasion, the time-scaled samples at 8kHz are assigned the same values as the rate converted samples at other sampling rates.

Figure 2.9 - Time-Scale Modification of CELP Coded Speech

stochastic excitation vector. In the 2:1 TSM, only half of the original stochastic vector is retained as the time-scaled vector.

In order to evaluate the two proposed methods, a small speech segment was time-scale modified (on a continuous basis) using each of the TSM schemes and the following results were obtained based on informal listening tests:

- For 4:3 and 4:5 TSM where playback rate was adjusted by $\pm 25\%$, a slight metallic distortion could be heard but the time-scaled speech was clearly intelligible and much of the speaker's characteristic was maintained when compared to the decoded speech at the nominal rate. Either TSM method yielded similar output speech quality.
- For 2:1 TSM where playback rate of output speech was doubled, the distortion became hardly noticeable since the synthetic speech was played back quite fast.
- For 1:2 TSM where playback rate of output speech was halved, there was a marked difference in speech quality between the two methods. In case of TSM via sampling rate conversion, output speech had a lot of crackling noise while simple replication of the stochastic excitation vector yielded much better speech quality.

The overall speech quality was deemed satisfactory in all five TSMs. Figure 2.11 and Fig. 2.12(a)-(d) depict the original and time-scale modified speech segments obtained via the replication and truncation method.

Percentage of TSM	Total No. of TSM frames	1:2 TSM	4:5 TSM	4:3 TSM	2:1 TSM
1%	25	5	7	5	8
2%	50	13	9	19	9
5%	125	32	28	32	33
10%	249	75	49	54	71

Table 2.4 - Time-Scale Modification Statistics

Upon proving the applicability of the TSM methods to the ICELP coder, the adaptive buildout delay mechanism was simulated in the following manner. A SI bit was generated

(a) Sample Rate Decimator

(b) Sample Rate Interpolator

(c) Fractional Rate Converter

Figure 2.10 - Sampling Rate Conversion **Figure 2.11 - Original Synthetic CELP Speech**

Figure 2.12(a) - 4:5 Time-scale Expanded Speech

Figure 2.12(b) - 4:3 Time-scale Compressed Speech

Figure 2.12(c) - 1:2 Time-scale Expanded Speech

Figure 2.12(d) - 2:1 Time-scale Compressed Speech

for every CELP encoded speech frame according to the VAD method described in Section 2.3.2. Then a sequence of random (but valid) PL bit patterns was generated such that only 1:2 and 2:1 TSMs were performed in the silent frames while 4:3 and 4:5 TSMs were performed in the non-silent frames. The replication and truncation TSM method was employed because of its better overall performance than the sampling rate conversion as observed in the preliminary listening tests. Speech frames assigned with PL=(X) were left unmodified.

Based on the TSM statistics shown in Table 2.4 for 2483 speech frames (with combined male and female speakers), informal listening tests were performed by several test subjects and the following results were obtained:

- For 1% and 2% TSM, there was no noticeable difference between the original and time-scaled speech.
- At 5% TSM, an occasional blurring of words could be heard but the speech quality was still acceptable.
- At 10% TSM, the variation in playback rates of the time-scale modified speech was clearly noticeable and the output speech quality was unsatisfactory.

2.4 Speech Frame Reconstruction in Frame Relay

In previous packet voice experiments [62][27], it was shown that either skipping the lost speech frame or replacing the lost frame with silence can seriously degrade output speech quality. To minimize output speech distortion, the most recent correctly received frame can be used to substitute the lost frame. If the frame loss condition persists under severe traffic congestion, the repeated speech frames are then attenuated until the speech output is completely muted and comfort noise may be used to replace the lost frame. Other alternative speech frame reconstruction algorithms are described in Section 4.1.

Speech frame reconstruction in the frame relay network can benefit from the holding (or buffering) delay employed by the adaptive buildout delay mechanism at the packet voice receiver. Assuming the frame loss statistic has a uniform random distribution due to the presence of data traffic in the frame relay network [66], the lost frame parameters can either be extrapolated or interpolated based on the ones received in the past and possibly

in the future. Fig. 2.13 (a) and (b) depict the two scenarios of voice frame reconstruction. In the first scenario, the $(n+3)^{\text{rd}}$ and $(n+4)^{\text{th}}$ frames are lost and speech parameters in these two frames can be extrapolated from the ones in the $(n+2)^{\text{nd}}$ and $(n+1)^{\text{st}}$ frames. In the second scenario, only the $(n+3)^{\text{rd}}$ frame is lost and the adjacent frames have already been received, the lost frame parameters can be interpolated from the ones in the $(n+2)^{\text{nd}}$ and $(n+4)^{\text{th}}$ frames. The methods of parametric extrapolation and interpolation are discussed in Section 4.2.

2.5 Summary

In this chapter, adaptation of frame relay for packet voice communication was described. The frame relay protocol was first reviewed and possible problems in packet voice transmission were identified. The CELP speech coding algorithm was presented as a viable means to encode low-bit rate speech in the frame relay network.

To address the problem of variable packet delay, an adaptive buildout mechanism of [13] coupled with voice activity detection and time-scale modification was described. In due course, simple VAD and TSM algorithms were designed and simulated. It is found that up to 2% of total frames can be time-scale modified without significant degradation in speech quality and hence shows the feasibility of the adaptive buildout delay mechanism.

Another important issue for voice transmission in the frame relay network is the loss of frames due to network congestion. Voice frame reconstruction algorithms must be used to minimize output speech distortion. In the next chapter, it is shown that the frame loss problem can also occur in the digital cellular network due to adverse channel conditions.

- (a) **Extrapolation:** CELP frame parameters in the previous two frames are used to estimate the lost frame parameters. The extrapolation can be used to recover up to 2 frames of lost speech parameters as shown.

- (b) **Interpolation:** CELP frame parameters in the previous and future frames are used to estimate the lost frame parameters.

Figure 2.13 - Frame Reconstruction via (a) Extrapolation and (b) Interpolation

Chapter 3

Applications of Speech Coding (Part II): Digital Cellular Radio

3.1 Introduction

In the past decade, we have witnessed the rapid emergence of cellular radio and wireless technology. Digital cellular radio networks are being introduced to meet surging user demands. In particular, the second generation digital systems provide better reuse of carrier frequencies and more efficient bandwidth utilization. It was reported a 5-10 times capacity improvement could be achieved over the analog system [49]. The North American IS-54 system allows three voice channels to be transmitted over a 30 kHz FM link using Time Division Multiple Access (TDMA). The gross rate of transmission is 13 kb/s of which 7.95 kb/s is dedicated to speech coding. The European GSM system also employs TDMA to transmit compressed speech at a gross rate of 22.8 kb/s (13.2 kb/s for speech coding and 9.6 kb/s for channel coding). Development of half-rate IS-54 and GSM codecs are currently underway to further reduce the transmission rate and double the channel capacity.

Speech coding in the digital cellular network differs from the frame relay network in several aspects. In contrast to variable delay in the frame relay network, TDMA employed by the cellular network always incurs a fixed delay on each voice frame. In this scheme, several fixed-rate bit streams are time-division multiplexed onto a radio carrier for transmission from the land station to the mobile stations. In the opposite direction, transmission for the mobile stations to the land station is again time-division multiplexed

over another radio carrier. The TDMA time sequencing is synchronized to a common clock frequency.

The cellular radio environment is somewhat more complex than the frame relay network as it is affected by the radio frequency, vehicular speed, channel bandwidth and geographical location of the vehicle. In general, the received signal is subjected to both long-term and short-term fading effects. Long-term fading or shadowing is caused by the terrain (for example, mountains, rivers and forests) between the base station and vehicle. Short-term fading is caused by multipath reflections of local scatterers around the vehicle. Different propagation paths (as a result of signal reflection) can have different time delay and signal attenuation. If the delay spread is small compared to a single bit duration, amplitude flat fading will result. Conversely, large delay spread causes frequency selective fading. Figure 3.1 depicts the quasi-periodic fading envelope received at a moving vehicle due to both constructive and destructive interference. It should be noted that Doppler shift is also introduced as the vehicle travels at high speed. However, when the vehicle is stationary and the channel is static, the signal envelope can stay in deep fades on some occasions.

For high fading rates the error distribution is more random and for low fading rates it is more burst like. Since both the VSELP and RPE-LTP coding algorithms encode input speech signal in a frame-by-frame format, the fading effect can be characterized by the Frame Error Rate (FER) and the Bit Error Rate (BER). In this chapter, different methods to combat random bit and frame errors are discussed for both of these speech coding schemes.

3.2 Digital Cellular Network Architecture

There are three types of communication elements in the digital cellular network, namely mobile units (vehicles), base stations and mobile switching offices. A base station exchanges radio signals with the mobile units and in turn the base station communicates with the mobile switching office which connects the wireless information network with the PSTN. In general, each base station serves around 30-40 wireless terminals while the mobile switching office is capable of handling radio channel assignments for up to 100 base stations [23].

Figure 3.1 - A Typical Fading Envelope at 835 MHz for a speed of 100 km/h

Figure 3.2 - GSM Digital Cellular Network Architecture

The following functional enhancements are provided by the digital cellular network:

Roaming: This refers to the capability of serving subscribers in locations remote from their home service areas. Communication between the mobile switching office and customer databases can be established through dedicated signalling protocols. As seen by the block diagram (Fig. 3.2) of the GSM cellular network, the mobile switching center is connected to three separate databases. The equipment identity register authenticates the mobile unit and the home location whereas the visitor location registers contain the necessary information to deliver calls to roaming users [23]. It can be noted that the signalling interfaces (labelled as A-G and U_m) in between any two functional blocks are defined in the GSM standard to ensure equipment compatibility among different manufacturers. The interface between base station and mobile switching office is not defined in the IS-54 system.

Mobile Assisted Handoff: In the current analog system, the mobile switching office collects information from several base stations about the signal strength for a particular mobile unit. The switching office then decides when the mobile unit should switch over to a new radio channel at a different base station. In contrast, the digital mobile stations are designed to monitor the strength of signals from nearby base stations. The measurement results are reported back to the corresponding base station which decides if a handoff is necessary and communicates back to the switching office information of the preferred new base station [49]. The system controller (inside the switching office) is therefore relieved from much of the processing burden of hand-offs. The result is a faster hand-off time for mobile users.

3.2.1 Overview of IS-54 Radio Interface

IS-54 is a dual-mode system capable of supporting the existing analog Advanced Mobile Phone System (AMPS) as well as providing digital services to new subscribers for higher capacity. Each 30 kHz carrier band supports a 40 ms TDMA frame which is subdivided into six 6.67 ms time slots as shown in Fig. 3.3 [21]. For a full-rate voice channel at 13 kb/s, two time slots of the frame: {1,4}, {2,5} or {3,6} are employed. Half-rate transmission at 6.5 kb/s which utilizes only a single time slot per frame will be used to double user capacity. Each 324-bit time slot includes either 260 bits of encoded voice

Figure 3.3 - IS-54 Time Slot and Frame Structure

Figure 3.4 - GSM Time Slot and Frame Structure

information (DATA), 12 bits of Slow Associated Control Channel (SACCH) control signalling information, a 28-bit synchronization pattern (SYNC) and a 12-bit Digital Verification Colour Code (DVCC). Additionally, a 6-bit guard time interval (G), a 6-bit power ramp-up interval (R) and a 12-bit reserved field (RSVD) are found in a timeslot sent in the mobile to base direction

The SYNC field contains a known bit sequence (with good autocorrelation properties) which enables the receiver to establish bit synchronization and train an adaptive equalizer. A unique synchronization pattern is used for each of the six time slots such that the receiver can lock onto its assigned time slot within the TDMA frame. The DVCC code serves as a channel identifier to avoid the receiver from locking onto an interfering co-channel signal from a distant cell. Furthermore, this 8-bit DVCC is protected by a (12,8) Hamming code and used to identify up to 256 different base stations.

The SACCH is a signalling channel dedicated for call processing. The control data stream is protected by a rate 1/2 convolutional coder such that given an original 66-bit control message, a 132-bit error protected message will result. This information is interleaved over 12 time slots for transmission in the SACCH field. For the full-rate channel, a single message can be sent in 6 frames (240 ms) and results in an information rate of 280 b/s. This information rate is halved when the IS-54 operates in the half-rate channel mode. An in-band Fast Associated Control Channel (FACCH) has also been defined by the standard for rapid handoff. To minimize output speech distortion, a lost frame recovery method must be employed to estimate speech frames which were deleted to make room for the FACCH messages.

The IS-54 modulation method is $\pi/4$ -shifted differential quaternary phase shift keying with square root raised cosine filtering at the transmitter and receiver. Given a transmission rate of 48.6 kb/s and a channel spacing of 30kHz, the modulation efficiency is 1.62 b/s/Hz.

3.2.2 Overview of GSM Radio Interface

Similar to the IS-54, the GSM multiple access technique is a combination of frequency division with carriers spaced at 200 kHz and time division with 8 data frames [21]. Figure 3.4 shows the organization of the GSM frame structure. A 120 ms multiframe is therefore

subdivided into 26 frames each of which contains eight time slots. The 24 frames are for user data and 2 frames are reserved for SACCH where up to 16 control channels can be specified. In a particular time slot, there are 2×57 DATA bits carrying user information and 2 corresponding F bits which act as "frame stealing flags". The F bits are used to indicate whether the DATA bits contain speech or signalling information. A 26-bit equalizer training sequence is placed in between the two sets of F and DATA bits and the 3 all-zero tail bits (T) are used to delimit the start and end of a timeslot. The guard time interval (GUARD) of 8.25 bits ensures no overlap occurs at the base station for signal bursts transmitted from different mobile units. With 156.25 bits every 577 μ s, the transmission rate is 270.833 kb/s. The full-rate GSM implementation uses 24 user frames and 1 SACCH frame in the multiframe structure (and the second SACCH frame is not used presently). For future half-rate operation, each voice channel will occupy 12 user frames and a SACCH frame.

GSM signalling messages are usually sent over the SACCH. Forward error correction codes are applied to 184 signalling information bits to produce 456 bits which are interleaved and distributed over four time slots. This 456-bit message is sent over an interval of 480 ms, resulting in a bit rate of 950 b/s. Similar to the IS-54 system, FACCH is employed to provide rapid handoff. The FACCH is created by "stealing" frames from the speech traffic channel in which case the special "frame stealing flags" (F bits) are set in the control part of the burst. Once again a lost frame recovery method must be applied to bridge the gaps.

GSM uses the Gaussian minimum shift keying technique for data modulation. Given a multiframe rate of 270.833 kb/s operating with a 200 kHz carrier spacing, the modulation efficiency is rated at 1.35 b/s/Hz.

An important speech processing function called Discontinuous Transmission (DTX) is employed in the GSM to increase spectral efficiency and reduce power consumption. The GSM DTX function consists of two parts: a voice activity detector and a comfort noise generator. The detector relies on the use of an inverse filter whose coefficients were derived during noise-only periods. The filter is then used to attenuate noise present in the received signal. The filtered signal is compared to a threshold which is updated during the noise-only period. If the signal energy is below the threshold over a hangover period, then

a silent frame is indicated. The comfort noise parameters are computed within the RPE-LTP framework at the transmitter during speech pauses and sent as a special Silence Descriptor (SID) frame over the channel. The comfort noise is important to ensure abrupt changes in speech level (due to DTX) does not cause annoying noise modulation effects at the speech decoder.

3.3 Speech Coding for Digital Cellular Communications

In this section, the VSELP and RPE-LTP speech coders standardized for use in the IS-54 and GSM digital cellular systems are described. The associated error control mechanisms are presented in Section 3.3.3.

3.3.1 Vector Sum Excited Linear Prediction (VSELP)

The bit allocation of VSELP coding scheme is tabulated below [14]:

Parameter	Bits/subframe	Bits/frame
Reflection coefficients: k_i	-	38
Frame energy $R_q(0)$	-	5
Stochastic codebook Indices: I, H	7+7	56
Adaptive Codebook lags: M	7	28
GS-P0-P1 codes	8	32
TOTAL	29	159

Table 3.1 - 8 kb/s VSELP Coder Bit Allocation

The basic structure of the VSELP speech decoder is shown in Fig. 3.5. The VSELP speech coder utilizes two MSVQ 40-dimensional stochastic codebooks and an one-tap adaptive codebook to generate the residual signal to excite a 10th order all-pole LPC filter. Moreover, the two stochastic codebooks are structured to reduce computational complexity. Three separate codebook gains: β , γ_1 and γ_2 are employed to scale the corresponding excitation vectors before summing them to form the residual excitation. Compared to VSELP, the ICELP coder employs three ACB coefficients (or LTP gain terms), one stochastic excitation gain value and one frame gain value to adjust the energy

(Note: quantized parameters are used.)

Figure 3.5 - Block Diagram of VSELP Speech Decoder

Figure 3.6 - VSELP Basis Vector Structured Codebook

in the residual signal.

In VSELP, the LPC coefficients and an average energy term $R_q(0)$ are coded once per 20 ms frame. The lag, stochastic excitation indices and various gains are updated at the 5ms subframe rate. Intermediate LPC values (at the subframes) can be interpolated directly from the coefficients (transformed from k_i) from two adjacent frames. Unlike ICELP where the short-term LPC coefficients are quantized using MSVQ, individual reflection coefficients are quantized via the fast lattice (FLAT) algorithm [14] to give a 38-bit code. The gain terms: β , γ_1 and γ_2 are jointly vector-quantized to form an 8-bit GS-P0-P1 code. Each speech frame which has $40 \times 4 = 160$ samples is encoded in 159 bits to yield an aggregate transmission rate of 7.95 kb/s. The pitch prefilter and spectral postfilter which have similar forms to the VSELP adaptive codebook and short-term LPC filter may be used to enhance the perceptual quality of decoded speech [20].

To reduce memory storage and improve codebook search efficiency, the two stochastic codebooks are organized in a highly structured fashion as shown in Fig. 3.6. A stochastic excitation vector drawn from each codebook is essentially the linear sum of a defined set of 40-dimensional basis vectors. The element $v_{k,i}(n)$ of a stochastic excitation vector $v_{k,i}$ can be written as:

$$v_{k,i}(n) = \sum_{j=1}^N \theta_{i,j} c_{k,j}(n) \quad (3.1)$$

where $N (=7)$ is the number of basis vectors, subscripts k is 1 for codebook 1 and 2 codebook 2 and $i = 0, 1, \dots, 2^N - 1$ are the 128 codewords in each codebook. Referring to the earlier discussion on structured codebook design (page 19), $c_{k,j}$ is the j^{th} basis vector (where $j = 1, 2, \dots, N=7$) and $\theta_{i,j}$ is a sign multiplier such that if bit j of codeword $i \in \{1, H\}$ is set, then $\theta_{i,j} = +1$, otherwise $\theta_{i,j} = -1$. Note that if all the bits in codeword i are complemented, the new excitation vector is also negated. The basis vector representation also makes the codeword more robust to single bit error because an incorrect bit can only change the sign of one basis vector.

The codebook search strategies are similar to the ones discussed in Section 2.3.1.1. The VSELP adaptive codebook is a special case of (2.2) with a single-tap coefficient and

may be written in the following form:

$$\frac{1}{B_n(z)} = \frac{1}{1 - \beta z^{-\lfloor \frac{n+m}{m} \rfloor}} \quad (3.2)$$

where n is the sample index within a subframe such that $0 \leq n < K_c$ and m denotes a particular lag value. Note the equation above also accounts for the adaptive codebook definition given in (2.17) with $q=0$. Assuming γ_1 and γ_2 are zero, (2.15) is the objective function that needs to be minimized and is repeated here in a more convenient form:

$$\epsilon^{(m)} = y_1^T y_1 - \frac{(y_1^T \hat{\mathbf{d}}^{(m)})^2}{\hat{\mathbf{d}}^{(m)T} \hat{\mathbf{d}}^{(m)}} \quad (3.3)$$

where y_1 is the input speech minus the ZIR response of $H(z)$ as defined by (2.12) and $\hat{\mathbf{d}}^{(m)}$ is the adaptive codebook vector filtered by $H(z)$. The optimal lag, M , is found by maximizing the fractional term in (3.3). To reduce the search complexity, $\hat{\mathbf{d}}^{(m)}$ can be written as:

$$\hat{\mathbf{d}}^{(m)}(n) = \sum_{i=0}^{\lfloor \frac{n}{m} \rfloor} \hat{\mathbf{d}}_p^{(m)}(n - im) \quad (3.4)$$

where $\hat{\mathbf{d}}_p^{(m)}(n)$ is the partial convolution sum of $\hat{\mathbf{d}}(n)$ and $h(n)$ and is defined as:

$$\hat{\mathbf{d}}_p^{(m)}(n) = \sum_{k=0}^{\min(n, m-1)} \hat{\mathbf{d}}(k-m) h(n-k) \quad (3.5)$$

From the above, a recursive relation can be obtained for $\hat{\mathbf{d}}_p^{(m)}(n)$:

$$\begin{aligned} \hat{\mathbf{d}}_p^{(m)}(0) &= \hat{\mathbf{d}}(-m) h(0), & n &= 0 \\ \hat{\mathbf{d}}_p^{(m)}(n) &= \hat{\mathbf{d}}_p^{(m-1)}(n) + \hat{\mathbf{d}}(-m) h(n), & 1 \leq n \leq K_c - 1 \end{aligned} \quad (3.6)$$

A truncated impulse response of $h(n)$ is used in (3.6) for computational reduction and the energy term in the denominator of (3.3) can also be simplified accordingly [14] for the closed loop pitch search. Once the optimal lag, M , is determined, the corresponding

residual excitation $\hat{d}^{(M)}$ is then given by:

$$\hat{d}^{(M)}(n) = \hat{d}\left(n - \left\lfloor \frac{n+M}{M} \right\rfloor M\right), \quad 0 \leq n \leq K_c - 1 \quad (3.7)$$

Given the optimal lag M , codebook 1 is searched assuming $\gamma_2 = 0$ (and γ_1 and β are "floating"). After the codebook 1 parameters are found, codebook 2 can be searched assuming $\gamma_1 = 0$. Joint optimization and hence computation of the gains β , γ_1 and γ_2 are done after the selection of codebook indices. At each stage of the codebook search, each codevector candidate is orthogonalized (prior to the computation of the WMSE) to decouple the effect of previous excitation vectors. Since the codevector is a linear sum of the basis vectors, the orthogonalization can be applied instead to the filtered basis vectors themselves.

During the first codebook search, the codevector $v_{1,i}$ must be orthogonalized with respect to the codevector $\hat{d}^{(M)}$ drawn from the adaptive codebook. First of all, the filtered codevector can be defined as:

$$f_{1,i}(n) = v_{1,i}(n) \otimes h(n) = \sum_{j=1}^N \theta_{i,j} q_{1,j}(n) \quad (3.8)$$

where $q_{1,j}(n)$ is the ZSR of $H(z)$ to the basis vector element $c_{1,j}(n)$. Applying the Gram-Schmidt procedure [46], the N basis vectors can be orthogonalized with respect to $\hat{d}^{(M)}$ as follows:

$$q'_{1,j}(n) = q_{1,j}(n) - \frac{\Psi_j}{\Gamma} \cdot \hat{d}^{(M)}(n) \quad (3.9)$$

where the terms Ψ_j and Γ are defined respectively as:

$$\Gamma = \sum_{n=0}^{K_c-1} (\hat{d}^{(M)}(n))^2, \quad \Psi_j = \sum_{n=0}^{K_c-1} \hat{d}^{(M)}(n) q_{1,j}(n) \quad (3.10 \text{ a,b})$$

In this way, the orthogonalized filtered vector can be written as:

$$\mathbf{r}_{1,i} = \sum_{j=1}^N \theta_{i,j} q'_{1,j}(n) \quad (3.11)$$

and then (3.11) can be used to maximize the following objective function:

$$\frac{(\mathbf{f}_{1,i}^T \mathbf{y}_1)^2}{\mathbf{f}_{1,i}^T \mathbf{f}_{1,i}} = \frac{\left(\sum_{n=0}^{K_c-1} f'_{1,i}(n) y_1(n) \right)^2}{\sum_{n=0}^{K_c-1} (f'_{1,i}(n))^2} = \frac{C_{1,i}^2}{G_{1,i}} \quad (3.12)$$

The above expression is essentially the same as (2.21) except y_2 has been replaced by y_1 in the numerator because of the orthogonalization.

Orthogonalization for the second codebook search can be done in the same fashion as above. The filtered codevector $\mathbf{f}_{2,i}$ drawn from codebook 2 is orthogonalized using (3.9) and (3.10 a,b) and it can then be orthogonalized with respect $\mathbf{f}_{1,i}$ to yield:

$$\mathbf{f}_{2,i} = \sum_{j=1}^N \theta_{i,j} q'_{2,j}(n) \quad (3.13)$$

where $q'_{2,j}$ is a filtered codevector orthogonal to $\hat{\mathbf{d}}^{(M)}$ and $\mathbf{f}_{1,i}$. The search criteria is identical to (3.7) except $\mathbf{f}_{1,i}$ is replaced by $\mathbf{f}_{2,i}$.

The VSELP search process can be expediated by encoding the stochastic codeword in Gray code format. In other words, if the codewords u and i are adjacent to each other such that only a single bit position v differs, then $\theta_{u,v} = -\theta_{i,v}$ and $\theta_{u,j} = \theta_{i,j}$ for $j \neq v$. Through this special arrangement, it is possible to show the numerator $C_{k,i}$ and denominator $G_{k,i}$ in (3.12) can be computed recursively using the following relations [20]:

$$C_{k,u} = C_{k,i} + 2\theta_{u,v} \sum_{n=0}^{K_c-1} q'_{k,j}(n) y_1(n), \quad 1 \leq v \leq N \quad (3.14)$$

and

$$G_{k,u} = G_{k,i} + \sum_{j=1}^{v-1} \theta_{u,j} \theta_{u,v} D_{k,j,v} + \sum_{j=v+1}^M \theta_{u,j} \theta_{u,v} D_{k,v,j} \quad (3.15)$$

where N is the number of basis factors and the term $D_{k,i,j}$ is defined as:

$$D_{k,i,j} = 4 \sum_{n=0}^{K_c-1} q'_{k,i}(n) q'_{k,j}(n), \quad 1 \leq i \leq j \leq N \quad (3.16)$$

The complementary nature of the codewords also means that only half of the codevectors need to be evaluated. The sign of C_i determines whether the selected codeword or its complement should be used.

The three excitation gains β , γ_1 and γ_2 are jointly optimized by searching an 8-bit VQ codebook containing GS-P0-P1 transformed vectors. In VSELP, the residual signal $\hat{d}(n)$ is a sum of three excitation signals:

$$\hat{d}(n) = \beta \hat{d}^{(M)}(n) + \gamma_1 v_{1,I}(n) + \gamma_2 v_{2,H}(n), \quad 0 \leq n \leq K_c - 1 \quad (3.17)$$

where $\hat{d}^{(M)}(n)$ = optimal adaptive codebook vector element
 $v_{1,I}(n)$ = optimal stochastic codeword I from codebook 1
 $v_{2,H}(n)$ = optimal stochastic codeword H from codebook 2

The total energy of the residual signal $\hat{d}(n)$ can be expressed as:

$$RS = K_c R_q(0) \prod_{i=1}^P (1 - k_i^2) \quad (3.18)$$

where $R_q(0)$ is the average quantized energy across a speech frame and k_i is the i^{th} reflection coefficient of the short-term synthesis filter. The value RS is further adjusted by an "offset" GS to yield a total effective residual energy $R = GS \cdot RS$. Consequently, the energy (or gain) values of the three excitation vectors can be represented as fractions of the average speech frame energy.

If $P0$ and $P1$ are the fractional values representing the respective energy contribution of $\hat{d}^{(M)}$ and $v_{1,I}$ towards R , then they can be defined as:

$$P0 = \frac{\beta^2 |\hat{d}^{(M)}|^2}{R}, \quad P1 = \frac{\gamma_1^2 |v_{1,I}|^2}{R} \quad (3.19 \text{ a, b})$$

where the two values are constrained by $0 \leq P0 \leq 1$ and $P0 + P1 \leq 1$. It is possible to

express β , γ_1 and γ_2 in terms of parameters GS , $P0$ and $P1$:

$$\begin{aligned}\beta &= \sqrt{\frac{RS \cdot GS \cdot P0}{|\hat{\mathbf{d}}^{(M)}|^2}} \\ \gamma_1 &= \sqrt{\frac{RS \cdot GS \cdot P1}{|\mathbf{v}_{1,I}|^2}} \quad \gamma_2 = \sqrt{\frac{RS \cdot GS \cdot (1 - P0 - P1)}{|\mathbf{v}_{2,H}|^2}}\end{aligned}\quad (3.20 \text{ a, b, c})$$

In the case when either no suitable lag which yields positive correlation in $\mathbf{y}_1^T \hat{\mathbf{d}}^{(m)}$ of (3.3) is found or the current adaptive codebook state is all-zero, then β is set to 0.

At this point, the WMSE ϵ which is the same as (2.18) can be written in the following form:

$$\epsilon = \left| \mathbf{y}_1 - \beta \mathbf{H} \hat{\mathbf{d}}^{(M)} - \gamma_1 \mathbf{f}_{1,I} - \gamma_2 \mathbf{f}_{2,H} \right|^2 \quad (3.21)$$

To eliminate conversion between the gain values and the equivalent fractions, (3.21) can be expressed directly in terms of GS , $P0$ and $P1$ such that:

$$\begin{aligned}\epsilon &= \mathbf{y}_1^T \mathbf{y}_1 - a \sqrt{GS \cdot P1} - b \sqrt{GS \cdot P1} - c \sqrt{GS (1 - P0 - P1)} \\ &\quad + d \cdot GS \sqrt{P0 \cdot P1} + e \cdot GS \sqrt{P0 (1 - P0 - P1)} \\ &\quad + f \cdot GS \sqrt{P1 (1 - P0 - P1)} + g \cdot GS \cdot P0 + h \cdot GS \cdot P1 \\ &\quad + i \cdot GS (1 - P0 - P1)\end{aligned}\quad (3.22)$$

where variables a to i are functions of RS , energy of the three excitation vectors and cross-correlations between a vector pair from $\mathbf{H} \hat{\mathbf{d}}^{(M)}$, $\mathbf{f}_{1,I}$ and $\mathbf{f}_{2,H}$ [14]. The VQ GS-P0-P1 codebook is exhaustively searched and the codeword which minimizes (3.22) is selected for encoding.

As indicated in (3.20a-c), the three gain terms are all normalized with respect to the energy in the residual excitation vector. Hence in situations where the GS-P0-P1 code is corrupted due to bit errors but the value of $R_q(0)$ is correctly received, the decoded speech frame will not result in sudden energy increase.

3.3.2 Regular Pulse Excited Long-Term Prediction (RPE-LTP)

Parameter	Bits/subframe	Bits/frame
LPC coefficients: LAR_i	-	36
LTP coefficient: b_q	2	8
LTP lag: M	7	28
RPE grids: N	2	8
block maxima: $x_{\max}$	6	24
RPE samples: x_N	39	156
TOTAL:	56	260

Table 3.2 - 13 kb/s RPE-LTP Bit Allocation

The functional block diagram of the RPE-LTP speech decoder and the bit allocations of the various parameters are shown in Fig. 3.7 and Table 3.2 respectively [16].

Similar to the VSELP design, the RPE-LTP utilizes a 20 ms frame for LPC analysis. Each frame is further divided into four subframes each of 5 ms for updating the long term predictor and the regular pulse excitation parameters. Eight reflection coefficients (k_i) are computed and converted into Log Area Ratios (LARs) which are defined as:

$$LAR_i = \log \left[\frac{1 + k_i}{1 - k_i} \right], \quad 1 \leq i \leq 8 \quad (3.23)$$

During the analysis stage, the LARs are transformed back into reflection coefficients and applied to an FIR lattice filter structure to analyze the original speech. Since the LARs are updated once per frame, the LAR values are interpolated at the subframe boundary. A single-tap long term predictor is used to remove long-term correlation in the speech signal. A cross-correlation lag search is performed on the residual excitation $d(n)$:

$$r(m) = \sum_{i=0}^{K_c-1} d(j \cdot K_c + i) \hat{d}(j \cdot K_c + i - m), \quad \begin{array}{l} j = 0, 1, 2, 3 \\ m = 40, \dots, 120 \end{array} \quad (3.24)$$

where $\hat{d}(n)$ is the combined signal from the regular pulse excitation $\hat{e}(n)$ and the long-term prediction defined as $\hat{d}(n) = b_q \hat{d}(n - M) + \hat{e}(n)$. The optimal lag M is found

Figure 3.7 - Block Diagram of RPE-LTP Speech Decoder

when $r(M) = \max[r(m)]$ for $m = 40, \dots, 120$. The LTP gain b of the j^{th} subframe is then calculated as:

$$b = \frac{r(M)}{\sum_{i=0}^{K_c-1} \hat{d}^2(j \cdot K_c + i - M)} \quad (3.25)$$

The stochastic excitation $e(n)$ after long-term prediction is weighted by an 11-tap FIR filter $h(n)$. The K_c central samples of the convolution sum: $x(n) = e(n) \otimes h(n)$ are retained and are down-sampled by a factor of three to 4 sub-sequences $x_k(n)$ each of length 13:

$$x_k(i) = x(j \cdot K_c + k + 3i), \quad \begin{matrix} i = 0, \dots, 12 \\ k = 0, \dots, 3 \end{matrix} \quad (3.26)$$

where k denotes the RPE grid candidate. The optimal sequence $x_N(n)$ is selected based on maximizing the total power of the excitation signal such that:

$$\epsilon_N = \max_k \left[\sum_{i=0}^{12} x_k^2(i) \right], \quad k = 0, \dots, 3 \quad (3.27)$$

Additionally, a block maxima $x_{\max}$ is obtained to normalize the regular pulse sequence. The 13 samples are then quantized using adaptive pulse coded modulation based on $x_{\max}$. In the decoder, the regular pulse excitation parameters N , $x_N(n)$ and $x_{\max}$ are decoded and the excitation sequence is reconstructed by placing two zero samples in between every $x_N(n)$ sample.

3.3.3 Error Protection Schemes

Since decoded speech deteriorates rapidly under adverse channel conditions, a suitable error control mechanism must be chosen to mitigate the channel errors. This section summarizes the IS-54 and GSM error control methods which include parameter classification, Cyclic Redundancy Check (CRC), channel coding and bad frame masking.

3.3.3.1 Bit Error Control Methods

The VSELP encoder first sorts the bits of the encoded speech parameters into three different classes according to their perceptual importance [54] (Table 3.3). The Class A

bits which are usually the MSBs of the speech parameters represent the most important perceptual information as any single bit error in this class can degrade the decoded speech significantly. These bits are protected in a two-stage fashion employing CRC and convolution coding. The Class B bits are the second group of perceptually important information. Bit errors in this class can affect the speech quality but not the intelligibility and hence these bits are protected only with convolutional coding. The Class C bits are the least important perceptually as bit errors usually result in noisier sounding speech. Consequently, the Class C bits are transmitted with no error protection.

VSELP Parameters	Bits/ Frame	Class A Bits	Class B Bits	Class C Bits
$R_q(0)$	5	3	1	1
$k_1 - k_{10}$	38	9	4	25
Lags	28	-	28	-
I, H indices	56	-	-	56
GS-P0-P1 codes	32	-	32	-

Table 3.3 - VSELP Parameter Class Bit Assignment

As discussed earlier, the 12 Class A bits are protected by a 7-bit CRC generated via the polynomial: $g_{\text{CRC}}(D) = 1 + D + D^2 + D^4 + D^5 + D^7$. The 19-bit resultant and the other 65 class B bits are then scrambled and appended with 5 zero tail bits before feeding into a convolution channel coder with half-rate and constraint length 5. As indicated in Fig. 3.8(a), the convolution code is the alternated output of two polynomials $g_0(D)$ and $g_1(D)$ defined as:

$$\begin{aligned} g_0(D) &= 1 + D + D^3 + D^5 \\ g_1(D) &= 1 + D^2 + D^3 + D^4 + D^5 \end{aligned} \quad (3.28)$$

The remaining 82 Class C bits are scrambled with the 178 convolution coded bits to form a 260-bit frame. A bit interleaving strategy which randomizes channel errors and minimizes total frame loss is employed before actual data transmission. The n^{th} frame is interleaved with the $(n-1)^{\text{st}}$ and $(n+1)^{\text{st}}$ frames over two timeslots. In the first time slot, the odd-bits (bit 1,3,...., 257, 259) of the n^{th} frame is transmitted with the even bits (bit 0,2,....,

256, 258) of the $(n-1)^{\text{st}}$ frame and in the second time slot, the even-bits of the n^{th} frame are transmitted with the odd-bits of the $(n+1)^{\text{st}}$ frame.

RPE-LTP Parameters	Bits/ Frame	Class Ia Bits	Class Ib Bits	Class II Bits
LAR_i	36	14	9	13
LTP lag	28	24	4	-
LTP gain	8	-	8	-
$x_{\max}$	24	12	8	4
N	8	-	8	-
x_N	156	-	95	61

Table 3.4 - RPE-LTP Parameter Class Bit Assignment

Similar error protection methods are used for the GSM RPE-LTP coder. An encoded GSM speech frame consists of: short-term LPC parameters, LTP parameters and RPE parameters. These three sets of parameters are divided into three different classes: (Ia, Ib and II) according to their perceptual importance. The bit classification is tabulated in Table 3.4.

The 50 Class Ia bits are protected by a 3-bit CRC code generated via the polynomial: $g_{\text{CRC}}(D) = 1 + D + D^3$. The 53 resulting bits are reordered and appended with 4 zero tail-bits. The 57 bits and 132 class Ib bits are then passed through a convolution encoder with half rate and constraint length 4 to result in $2 \times 189 = 378$ bits. Referring to Fig. 3.8(b), the GSM convolutional codes are defined by the two polynomials [58]:

$$\begin{aligned} g_0(D) &= 1 + D^3 + D^4 \\ g_1(D) &= 1 + D + D^3 + D^4 \end{aligned} \quad (3.29)$$

The remaining 78 class II bits are left unprotected. The combined convolution coded bits and class II bits (456 bits in total) are again reordered. Bit and frame interleaving are applied prior to transmission. The 456 bits of the n^{th} frame are distributed over the 8 timeslots of a frame shown in Fig. 3.9 [26]. For timeslots 1-4, the n^{th} frame bits are interleaved with the $(n-1)^{\text{st}}$ frame bits such that the n^{th} frame bits always occupy the even-

(a) VSELP Convolution Coder

(b) RPE-LTP Convolution Coder

Figure 3.8 - VSELP and RPE-LTP Channel Coders

Key: (a) Interleaving
 (b) Adding Stealing Flags
 (c) Final Mapping

Figure 3.9 - Interleaving GSM Speech Frames

bit positions while the $(n-1)^{\text{st}}$ frame bits occupy the odd-bit positions within the timeslots. For timeslots 5-8, the n^{th} frame bits are interleaved with the $(n+1)^{\text{st}}$ frame bits. On this occasion, the odd and even bit positions are interchanged between the n^{th} frame bits and the $(n+1)^{\text{st}}$ frame bits. For each time-slot, there are 57 bit contributions from both the current and the next encoded speech frames and these are appended with two frame "stealing flags" to result in 116 bits. Additionally, the next encoded frame starts at every 4th timeslot and its bits are distributed over 8 timeslots.

At the receiver side, the IS-54 and GSM systems undo the bit interleaving. The convolution encoded data can be decoded using the Maximum Likelihood Sequence Estimation (MLSE) via the Viterbi algorithm [46]. At this stage, the decoder attempts to correct any random bit errors in the data stream. Furthermore, errors in the perceptually important bits: VSELP class A bits and GSM class 1a bits are detected by comparing the received CRC with the one generated locally at the receiver. If there is a mismatch, then an error has occurred in the perceptually important bits. In this case, the corrupted parameters are discarded and appropriate bad frame masking procedure is invoked to reduce output speech distortion. In the next section, the lost frame recovery methods employed by either system are reviewed.

3.3.3.2 Frame Error Recovery Methods

Frame loss can arise in the digital cellular radio due to several reasons. Firstly, the speech frames may have to make way for in-band control signalling data and secondly, frames can be lost due to fading conditions. When channel errors are random or could be randomized by bit interleaving, Forward Error Correction (FEC) methods could be used effectively. However, the channel can become either become very good or very bad for a prolonged period of time under certain circumstances (for example when the vehicle enters a tunnel.) Even though bit interleaving is applied, the error burst can no longer be randomized due to the high error probability which can approach 0.5 in the worst case [26]. Convolutional coding will fail to correct the bit errors and lead to bad CRC for the perceptually important parameters. Increasing the amount of bit interleaving could eventually randomize the errors but would result in unacceptable delay.

The VSELP decoder has a bad frame masking strategy [14] based on a 6-state finite state machine. The state number identifies number of consecutive frames having bad

CRC. In state 0, there is no loss in the last frame and in state 6, at least 6 previous consecutive frames have been lost. The state machine may change its state depending upon the current error condition. Every time a bad CRC is encountered, the state count increments by one until the state count reaches 6 where the state simply repeats itself again. If the current state number is between 1 and 5 and the received parameters have correct CRC, the FSM will return to state 0. State 6 requires at least two contiguous frames with correct CRC to return to state 0. This is to prevent false indication of good CRC under poor channel conditions. The corresponding corrective actions at each state are summarized below:

- State 0:** Received CRC is correct, no action is required, the decoded parameters are sent to the VSELP decoder.
- State 1:** A CRC error has been received; the receiver repeats the frame energy $R_q(0)$ and reflection coefficients k_i of the last frame in state 0. The other decoded bits for stochastic codebook indices, lags, GS-P0-P1 codes are sent to the speech decoder without modification.
- State 2:** Same as state 1.
- State 3:** Same as state 1 and 2 except $R_q(0)$ uses a 4 dB less version of the frame energy of the last correctly received frame.
- State 4:** Same as state 3 except $R_q(0)$ uses a 8 dB less version of the frame energy of the last correctly received frame.
- State 5:** Same as state 4 except $R_q(0)$ is attenuated 12dB.
- State 6:** Frame is repeated but $R_q(0)$ is set to 0 thereby removing the residual excitation. The speech output will only be the ZIR of the short-term synthesis filter and will gradually decay to silence. Comfort noise may be inserted to replace the entire frame.

A recent enhancement to the VSELP bad frame masking procedure was reported in [54]. In addition to the above, a Hamming distance is used to compute a maximum-likelihood metric during MLSE channel decoding in the VSELP. If the resultant metric exceeds the preset threshold, then the lag and subframe gain parameters are also copied from the last subframe of the previous frame correctly received. It was shown that copied values helped to improve the reconstruction fidelity under heavy fading conditions.

The GSM standard recommends lost frame substitution methods based on either repetition or extrapolation of the previous good speech frame(s). In the first approach, the receiver replaces the parameters of the first lost frame with the ones from the last correct speech frame [26][17]. For subsequent lost speech frames, the decoded speech level is attenuated gradually by increasing the value of $x_{\max}$ such that complete silence is obtained after 320 ms. The RPE grid positions can be chosen randomly during these lost frames. If the speech loss condition persists, then predefined "silence frames" are passed to the decoder to ensure low output level under all conditions. Treatment of lost SID frames during DTX is identical. The alternative frame substitution method based on speech extrapolation (or prediction) is not specified by the current standard. In [18], speech extrapolation based on Mean Squared Error (MSE) and Maximum A Posteriori (MAP) estimator are described for use in the GSM RPE-LTP decoder. In the latter approach, the lost parameters are estimated by maximizing the following *a posteriori* probability:

$$P(\mathbf{a}/\mathbf{r}) = \frac{P(\mathbf{r}/\mathbf{a})P(\mathbf{a})}{P(\mathbf{r})} \quad (3.30)$$

where $\mathbf{a}$ and $\mathbf{r}$ are the transmitted and received signal parameter vectors. Equation (3.30) makes use of the *a priori* probability $P(\mathbf{a})$ of the encoded speech parameters. It was found that the MSE estimator has better perceptual performance than the MAP estimator [18].

3.4 Summary

In this chapter, the IS-54 and GSM digital cellular standards were reviewed. The VSELP and RPE-LTP speech coding schemes were then described as two efficient means for voice transmission over the mobile radio channel. Robust coding techniques such as Gray coding of stochastic codebook indices and gain value normalization were explained in the VSELP design. In order to cope with extreme conditions in the mobile radio channel, error control mechanisms such as CRC, channel coding and bit interleaving were introduced to mitigate the effect of random bit errors. For heavy fading conditions, the entire speech frame can be lost and bad frame masking is employed to reduce output speech distortion. In the VSELP decoder, selected parameters of the most recent correctly received frame are repeated for the lost frame. On the other hand, the RPE-LTP bad frame masking method repeats all of the previous parameters.

Although the study of random bit errors has received a lot of attention, work on correcting burst errors in the speech frame is mostly limited to the use of last frame repetition described above. Improvements could be made in handling frame loss where major degradation in speech quality is expected. In the next chapter, the subject of voice frame reconstruction is discussed. Vector linear prediction techniques suitable for use in CELP speech coders will be examined in detail.

Chapter 4

Reconstruction of Missing Speech

4.1 Literature Review

The requirement of reconstruction of missing speech stems from the packet loss problem in packet switched networks [27][65]. Although the reconstruction methods to be discussed vary according to the speech coding schemes, all of them strive to sustain the flow of speech by filling in the lost packets[†] rather than skipping them to minimize output speech distortion.

Much of the early work in this area was related to packet reconstruction methods for PCM and ADPCM speech. In [28], a “sample interleaving” strategy was employed whereby alternate PCM samples were packed separately in two packets. When one of two packets was lost, an estimate could be made by interpolating the speech samples. In [62], several packet reconstruction methods were considered, including: (1) silence fill-in, (2) last packet repetition, and (3) filling the gap with repeated frames of speech data, making the speech voiceless and decreasing speech energy exponentially. Whereas replacing lost packets with silence has the worse perceptual effect, the third method was found to work best, especially in the case of the LPC vocoder. In another investigation of ADPCM speech reconstruction, the lost packets were replaced by silence and then the first sample of the following packets was set to zero [28]. It was noted that speech degradation became noticeable at packet loss rates above 2%.

[†] Unless otherwise stated, a speech packet is equivalent to a single speech frame. The terms “packet” and “frame” are used interchangeably in this chapter.

In the next few sections, a number of established packet reconstruction methods are examined for PCM, ADPCM, APC and CELP coded speech. As it will be seen later (Section 4.2), several basic principles originally developed for PCM speech can be extended for use in the low bit rate CELP coding algorithm.

4.1.1 Pattern Matching Method

Relying on the short-term stationary characteristics of speech, waveform substitution methods based on pattern matching were developed in [27][24][61]. In the 2-sided version shown in Fig. 4.1, two blocks of M speech samples which are on either side of the lost packet are chosen as matching templates. A suitable pattern matching algorithm is then employed to identify speech segments that best match the templates. These speech segments are drawn from two search windows (of length N) located $L+P$ samples away from either side of the lost packet. The value L denotes the number of missing speech samples and P denotes the merge length. In case of multiple packet losses, the search windows are placed $2L+P$ samples away and if the packet is still missing within the future search window (of the 2-sided scheme), the packet is treated as silence and ignored by the search.

The choice of pattern matching method was not critical in the reconstructed speech quality [27]. Hence either one of the following methods: waveform cross-correlation $C(n)$, sign cross-correlation $S(n)$ or normalized magnitude difference $D(n)$ can be used:

$$\begin{aligned}
 C(n) &= \frac{\mathbf{x} \cdot \mathbf{y}_n}{|\mathbf{y}_n|^2} \\
 S(n) &= \sum_{i=1}^M \text{sgn}[x(i)] \text{sgn}[y(n+i)] \\
 D(n) &= \sum_{i=1}^M \left| \frac{x(i)}{|\mathbf{x}|} - \frac{y(n+i)}{|\mathbf{y}_n|} \right|
 \end{aligned} \tag{4.1a,b,c}$$

where n is the sample index of the search window; $\mathbf{x}$ and $x(i)$ denotes the M dimensional template vector and template sample and $\mathbf{y}_n$ and $y(i)$ are the candidate vector and candidate sample.

By sliding the template along the search windows and applying an appropriate pattern

Figure 4.1 - Waveform Recovery via Pattern Matching Method

matching algorithm, two $L+P$ speech segments: one segment immediately after the best matched pattern in the previous search window and another segment immediately before the best matched pattern in the future search window are chosen. These are combined linearly to form a single packet estimate. A simpler 1-sided approach can also be used involving only the previous search window.

In the 2-sided approach, the two best-matched segments are combined linearly by placing more weights on the previous segment than on the future segment. Hence the $L+2P$ samples from the previous search window are weighted with the scale factor $W_p(i)$ defined as:

$$W_p(i) = \frac{L+2P+i}{L+2P-1}, \quad i = 1, 2, \dots, 2P+L \quad (4.2)$$

and the weights for future $L+2P$ samples are:

$$W_f(i) = 1 - W_p(i) \quad (4.3)$$

The two sets of $L+2P$ samples are multiplied by the above weights and added together. To ensure there is no energy increase in the reconstructed packet when compared to previous speech, amplitude normalization based on either square-root of speech energy or sum of absolute magnitudes are used to scale the reconstructed packet to have signal amplitude no larger than that of the previous speech. It was reported proper normalization can significantly enhance the output speech quality, a fact which will be of use in the later development of an adaptive energy control mechanism for the CELP coded speech.

The normalized packet estimate is blended with the adjacent speech to ensure there is no waveform discontinuity. A raised cosine weighting function which yields satisfactory result was reported in [27]. The last P samples of the previous packet are assigned the following weights:

$$W_1(i) = \frac{1}{2} \left[1 + \cos \left[\frac{\pi(i-1)}{P-1} \right] \right], \quad i = 1, 2, \dots, P \quad (4.4)$$

and the first P samples of the packet estimate are weighted as follows:

$$W_2(i) = 1 - W_1(i) \quad (4.5)$$

Conversely, the last P samples of the packet estimate is weighted by $W_1(i)$ and the first P samples of the future packet is weighted by $W_2(i)$. The merge segments can be added together to result in continuous boundaries across the reconstructed and adjacent packets.

4.1.2 Pitch Detection Method

In the pitch detection method, a simple center-clipped peak detector could be employed to obtain a pitch period M if the lost packet is voiced [24]. The algorithm simply repeats the last M samples of the previous speech for the length of the missing packet. If the lost packet is unvoiced, then the algorithm simply repeats the previous packet. Merging is necessary to provide speech continuity at the boundaries between normal and reconstructed speech.

An enhanced pitch detection method utilizing phase-matching was reported in [59]. Although a smooth transition can be guaranteed at the beginning of the reconstructed packet, discontinuity can occur at the end of the packet (if the lost packet length is not a multiple of the pitch period.) The phase matching method essentially modifies the time-scale of the waveform to accommodate the time-variation of the pitch and waveform amplitude. As a result, the ending phase of the missing packet always matches the beginning phase of the following packet.

If the phase of the i^{th} sample of lost packet j is defined to be the index of the pitch period M :

$$\theta_j(i) = i \oplus M, \quad i = 1, 2, \dots, N \quad (4.6)$$

where N is the packet size. Then the phase discontinuity can be corrected either via waveform expansion or waveform compression by a distance d defined as:

$$d = \min(d_p, d_c) \quad (4.7)$$

where d_c and d_e are the compressed and expanded distances (see Fig. 4.2):

$$d_c = \begin{cases} \theta_j(N) - \theta_{j+1}(1), & \theta_j(N) \geq \theta_{j+1}(1) \\ \theta_j(N) - \theta_{j+1}(1) + M, & \text{otherwise.} \end{cases} \quad (4.8)$$

Figure 4.2 - Phase Correction in Pitch Detection Method

$$d_e = M - d_c \quad (4.9)$$

A time-scale modification factor α which depends on d_e and d_c can be derived as below:

$$\alpha = \begin{cases} \frac{N - d_e}{N}, & d_e \leq d_c \\ \frac{N + d_c}{N}, & d_e > d_c \end{cases} \quad (4.10)$$

The phase matched sequence is then linearly interpolated such that the phase-matched reconstructed samples can be written as:

$$x_j(\alpha n) = [x_j(\lceil \alpha n \rceil) - x_j(\lfloor \alpha n \rfloor)] (\alpha n - \lfloor \alpha n \rfloor) + x_j(\lfloor \alpha n \rfloor) \quad (4.11)$$

where $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ denote the ceiling and floor functions. Note that the above equation defines $x_j(\alpha n)$ for integral αn values only. A linear weighting function is further applied to remove waveform discontinuity:

$$w(n) = 1 + \frac{A_{j+1} - A_{j-1}}{A_{j-1}} \cdot \frac{n}{N} \quad (4.12)$$

where A_j is either the positive or negative peak amplitude value depending upon whether the value of $x_j(\alpha n)$ is positive or negative.

4.1.3 Double Sided Periodic Substitution (DSPS) Method

The DSPS is a simple variant of the PCM waveform substitution method [56]. Based on the observation that the lost speech packet should have close resemblance to speech packets on either side of the lost packet, an iterative substitution mechanism can be used to fill the speech gap. Defining the number of lost packets to be $m-1$ and the speech packets before and after the lost packet as two N -dimensional vectors $\mathbf{x}_0$ and $\mathbf{x}_m$, the missing packets are replaced by scaled versions of $\mathbf{x}_0$ and $\mathbf{x}_m$ starting from either end until the two packet replacements merge at the k^{th} lost packet (where $k \leq m$).

The DSPS algorithm can be defined according to the following set of equations:

$$\mathbf{x}_i = \begin{cases} \frac{\delta(i)}{\delta_0} \mathbf{x}_0, & 1 \leq i < k \\ \frac{\delta(k)}{\delta_0} \mathbf{x}_k, & i = k \\ \frac{\delta(i)}{\delta_m} \mathbf{x}_m, & k < i < m \end{cases} \quad (4.13)$$

where δ_0 and δ_m are the standard deviations of vectors $\mathbf{x}_0$ and $\mathbf{x}_m$ and $\delta(i)$ represents the linearly interpolated values between δ_0 and δ_m and is given by:

$$\delta(i) = \delta_0 + \frac{\delta_m - \delta_0}{m} \cdot i \quad (4.14)$$

and $\mathbf{x}_k$ is the merged packet defined as a linear combination of $\mathbf{x}_0$ and $\mathbf{x}_m$:

$$x_k(n) = \frac{N-n}{N-1} x_0(n) + \left[1 - \frac{N-n}{N-1} \right] x_m(n) \quad (4.15)$$

4.1.4 Packet Reconstruction for ADPCM and APC

The three prior methods are useful for reconstructing PCM speech and can tolerate random loss rates up to 10% for packet sizes 8 - 16ms [61] without significant degradation in the decoded speech quality. This is not necessarily the case when the same methods are applied to adaptive coding schemes such as ADPCM and APC. The problem in this case is that although the decoded speech waveform can be reconstructed using one of these schemes, the errors still remain in the coder memory and they are left uncorrected. Consequently, a propagation effect can result and cause speech distortion in the subsequent frames.

In [27], it was found that reconstruction on the ADPCM parameters performs poorly. The main reason is that backward adaptation in the decoder causes rapid divergence of the quantizer step-size and predictor coefficients from its original due to the memory hangover. A better reconstruction approach is to freeze the decoder state by not feeding it

any input data. At the PCM output, either PCM speech from the previous packet or PCM speech based on waveform substitution can be played back. When the next correctly received packet arrives, the ADPCM decoder can be updated as usual. Alternatively, an ADPCM encoder can be employed at the receiver side to encode the waveform substituted speech into the corresponding ADPCM parameters for updating the ADPCM states during the missing interval. The former reconstruction method of freezing the ADPCM parameter was reported to be perceptually better [27].

A packet recovery method based on embedded coding was first described for DPCM in [22] and further extended for ADPCM in [55][31]. As illustrated in Fig. 4.3, the idea is to employ a noisy version of the codeword (which has m LSBs deleted from the original N -bit codeword) for the prediction of the next sample in the feedback loop. At the decoder, m LSBs are again deleted from the received codeword for reconstruction purposes. The $(N-m)$ MSBs and m LSBs are transmitted in different packets and a packet prioritization protocol is then utilized to drop the packets containing the LSBs during heavy traffic congestion. The decoder state is the same regardless of dropping the low priority packets.

An “analysis-by-synthesis” embedded coding scheme has been proposed for an APC speech coder [38]. A packet predictor based on the pitch detection method is introduced into the sample predictor loop as illustrated in Fig. 4.4. During the encoding stage, the packet predictor will generate an estimate of reconstructed speech packet and if this estimation is judged to be close to the actual reconstructed packet (via cross-correlation), a packet substitution will occur and the predicted version is then used to provide the next sample prediction. The codeword (at the adaptive quantizer output) is marked as “non-essential”. On the other hand, if the packet substitution is unsuccessful, the transmitted codeword is marked as “essential”. At the receiver, packet substitution occurs normally for non-essential packets. A packet prioritization protocol is used to allow selective packet droppings. At the receiver (Fig. 4.4), the decoded speech is normally obtained from output A except during a packet loss when the decoded speech is obtained from output B.

4.1.5 Packet Reconstruction for CELP

Many of the aforementioned packet reconstruction methods are equally applicable for CELP speech coders. In [32][65][10], a waveform substitution approach is used to

Figure 4.3 - Embedded Coding System

Figure 4.4 - Embedded APC Speech Coder

reconstruct the lost frames. If the lost packet is estimated to be unvoiced, then the output can be reconstructed by repeating the last frame parameters. Alternatively if it is voiced, then the output is generated by repetitions of scaled versions of the previous period (just before the missing frame.) In the parametric domain, this involves repeating the lag (of the last subframe) and scaling the previous codebook gain. No waveform merging is necessary because the reconstruction takes place in the parametric domain.

In [65], two prioritized marking schemes were proposed for CELP coders. The first scheme uses a packet classifier to distinguish between low and high priority packets. The low priority packets include low energy packets as well as packets which show little variation from the previous ones. High priority packets include the onset speech packets and highly transitional packets. The classification is obtained by examining the speech energy, pitch predictor coefficient and the log-spectral distance of each packet. In the second prioritization scheme, the CELP parameters are split into separate packets and only the low-priority packets are dropped during traffic congestion.

Embedded coding has also been applied to CELP coders. A 2-stage embedded 16 kb/s CELP coder [65] can be used. Speech is coded initially at 8 kb/s in the first stage. Then the residual between the synthetic speech (as a result of first stage decoding) and the original speech is encoded by a second stage coder to yield an 8 kb/s auxiliary data stream. During normal decoding, the two data streams are decoded separately and summed together to form decoded speech at twice the rate. In another embedded CELP coder design, the auxiliary 8 kb/s data is specified by a multi-pulse excitation algorithm [37]. The auxiliary information can be dropped with the quality degraded to that of the embedded 8 kb/s coder.

It should be noted that although embedded coding is applicable in packet networks with prioritized voice service, it is less useful in the TDMA digital cellular applications discussed in Chapter 3 unless some sort of dynamic channel allocation is employed. It is equally likely "essential" and "non-essential" packets will be dropped.

4.2 Speech Frame Extrapolation and Interpolation

In this section, alternative packet reconstruction algorithms for CELP speech coders are described based on parametric extrapolation and interpolation. Vector Linear

Prediction (VLP) which is viewed as the optimal approach for linear extrapolation and interpolation is reviewed.

4.2.1 Vector Linear Prediction (VLP)

The subject of Linear Prediction (LP) can be viewed as a special case of least-squares estimation or linear regression theory. In essence, LP is concerned with the estimation of some "future" random variables using a linear combination of a set of random variables observed in the "past". In mathematical terms, if we let $\mathbf{X}_j$ be a k -dimensional vector with finite energy, $E[\|\mathbf{X}_n\|^2] < \infty$, drawn from a stationary random sequence of N observed vectors with zero mean such that:

$$\mathbf{X}_j = (x_{j1}, x_{j2}, \dots, x_{jk})^T, \quad 1 \leq j \leq N \quad (4.16)$$

where x_{jn} is the i^{th} element, it is possible to predict another m dimensional random vector $\mathbf{Y}$ belonging to the same stationary process:

$$\mathbf{Y} = - \sum_{j=1}^N \mathbf{A}_j \mathbf{X}_j \quad (4.17)$$

where $\{-\mathbf{A}_j\}$ is a set of N $m \times k$ LP coefficient matrices (Note: m and k are not necessarily the same). The above thus defines the general form of a linear predictor with the underlying assumption $\mathbf{X}_j$ and $\mathbf{Y}$ being non-deterministic; $\mathbf{Y}$ cannot be predicted exactly from $\mathbf{X}_j$ with non-zero probability [19].

There are two cases which are of particular interests in this thesis. In the first case in which $m=k=1$, $\mathbf{X}_j$ and $\mathbf{Y}$ then take on scalar values $x_{n,j}$ and $\hat{x}_n$. Equation (4.17) becomes:

$$\hat{x}_n = - \sum_{j=1}^N a_j x_{n-j} \quad (4.18)$$

where $x_{n,j}$ are the N previous observations $x_{n-1}, x_{n-2}, \dots, x_{n-N}$ of a time series, $\hat{x}_n$ is an estimate of the present observation x_n and $\{a_j\}$ is the set of N scalar predictor coefficients. Equation (4.18) defines a *scalar linear predictor* which is used in the formulation of the short-term and long-term LP filters of the CELP speech coder in (2.1) and (2.2).

In the second case $m = k > 1$, $\mathbf{X}_j$ are N prior observations of the vector $\mathbf{X}_n$: $\mathbf{X}_{n-1}, \mathbf{X}_{n-2}, \dots$,

$\mathbf{X}_{n-N}$ and $\mathbf{Y}$ is an estimate of $\mathbf{X}_n$ such that $\mathbf{Y} = \hat{\mathbf{X}}_n$, (4.17) then becomes:

$$\hat{\mathbf{X}}_n = - \sum_{j=1}^N \mathbf{A}_j \mathbf{X}_{n-j} \quad (4.19)$$

where $\{-\mathbf{A}_j\}$ is a set of N $k \times k$ square matrices. Equation (4.19) defines a **vector linear predictor**. Internally, the vector predictor operates as k separate scalar predictors. If we let $\hat{x}_{ni}$ be the i^{th} element of $\hat{\mathbf{X}}_n$, and $\mathbf{a}_{ji}^T$ be the i^{th} row vector of the LP coefficient matrix $\mathbf{A}_j$, (4.19) can be rewritten as:

$$\hat{x}_{ni} = - \sum_{j=1}^N \mathbf{a}_{ji}^T \mathbf{X}_{n-j}, \quad i = 1, 2, \dots, k \quad (4.20)$$

It can be seen from (4.20) that each element of the vector estimate is predicted not only from corresponding element of previous vectors, but also from other elements of previous vectors. In this fashion, the correlation between all elements of the different vectors can be exploited.

In order to determine the optimal set of LP coefficient matrices $\{-\mathbf{A}_j\}$, an objective measure is needed and the MSE measure can be used:

$$D_{\text{MSE}} = E[\|\mathbf{X}_n - \hat{\mathbf{X}}_n\|^2] = E[\|\mathbf{e}_n\|^2] \quad (4.21)$$

where $\mathbf{e}_n$ is the prediction residual vector. Alternatively, another measure known as prediction gain can also be used as defined below:

$$G_{p(\text{dB})} = 10 \cdot \log_{10} \left| \frac{E[\|\mathbf{X}_n\|^2]}{E[\|\mathbf{e}_n\|^2]} \right| \quad (4.22)$$

A vector linear predictor is said to be optimal if it either minimizes (4.21) or maximizes (4.22). In the special case of scalar linear prediction given by (4.18), the set of scalar predictive coefficients $\{a_j\}$ can be found by differentiating D_{MSE} with respect to each coefficient and setting the derivative equal to zero. This would yield the Wiener-Hopf equations which is a set of N linear equations with N unknowns and can be expressed conveniently in matrix form as:

$$\begin{bmatrix} r_0 & r_1 & \cdots & r_{N-1} \\ r_1 & r_0 & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & r_1 \\ r_{N-1} & \cdot & \cdots & r_1 & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_N \end{bmatrix} = - \begin{bmatrix} r_1 \\ r_2 \\ \cdot \\ \cdot \\ r_N \end{bmatrix} \quad (4.23)$$

where $r_i = E[x_n x_{n+i}^T]$ are the autocorrelation coefficients. It can be seen the autocorrelation matrix in (4.23) is both symmetric and Toeplitz.

The same calculus approach could also be used to solve (4.21) for the vector case but would lead to unwieldy mathematics. A more elegant way is to invoke the orthogonality principle [9], which states that the prediction residual vector $\mathbf{e}_n$ must be orthogonal to all of the observed random vectors $\mathbf{X}_j$. In this case, it is possible to write:

$$E[\mathbf{e}_n \mathbf{X}_{n-j}^T] = \mathbf{0}, \quad j = 1, 2, \dots, N \quad (4.24)$$

Since $\mathbf{e}_n = \mathbf{X}_n - \hat{\mathbf{X}}_n$, the following is obtained by substituting in (4.20) for $\hat{\mathbf{X}}_n$:

$$E\left[\left(\mathbf{X}_n + \sum_{i=1}^N \mathbf{A}_i \mathbf{X}_{n-i}\right) \mathbf{X}_{n-j}^T\right] = \mathbf{0} \quad (4.25)$$

which can be simplified to the following:

$$\mathbf{R}_{0j} = - \sum_{i=1}^N \mathbf{A}_i \mathbf{R}_{ij}, \quad j = 1, 2, \dots, N \quad (4.26)$$

The above equation can be further expressed as:

$$\begin{bmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} & \cdots & \cdot & \mathbf{R}_{1N} \\ \mathbf{R}_{21} & \mathbf{R}_{22} & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \mathbf{R}_{N-1N} \\ \mathbf{R}_{N1} & \cdot & \cdots & \mathbf{R}_{NN-1} & \mathbf{R}_{NN} \end{bmatrix} \begin{bmatrix} \mathbf{A}_1^T \\ \mathbf{A}_2^T \\ \cdot \\ \cdot \\ \mathbf{A}_N^T \end{bmatrix} = - \begin{bmatrix} \mathbf{R}_{10} \\ \mathbf{R}_{20} \\ \cdot \\ \cdot \\ \mathbf{R}_{N0} \end{bmatrix} \quad (4.27)$$

where $\mathbf{R}_{ij} = E[\mathbf{X}_{n-i} \mathbf{X}_{n-j}^T]$ and can be rewritten compactly using the following notations:

$$\mathbf{R}\mathbf{A} = -\mathbf{V} \quad (4.28)$$

where $\mathbf{R}$, $\mathbf{V}$ and $\mathbf{A}$ are the respective supermatrix, supervector and LP supervector as defined in (4.27). Comparing (4.27) with (4.23), it is obvious (4.27) is the vector version of Wiener-Hopf Equations but unlike its scalar counterpart, the autocorrelation supermatrix $\mathbf{A}$ can only be described as symmetric but, in general, non-Toeplitz.

Returning to (4.21), the MSE of the vector linear predictor can be expressed in terms of $\mathbf{R}$, $\mathbf{V}$ and $\mathbf{A}$ to yield:

$$\begin{aligned} D_{\text{MSE}} &= E \left[\left(\mathbf{X}_n + \sum_{i=1}^N \mathbf{A}_i \mathbf{X}_{n-i} \right)^T \left(\mathbf{X}_n + \sum_{j=1}^N \mathbf{A}_j \mathbf{X}_{n-j} \right) \right] \\ &= E[\mathbf{X}_n^T \mathbf{X}_n] + 2E \left[\mathbf{X}_n^T \sum_{i=1}^N \mathbf{A}_i \mathbf{X}_{n-i} \right] + E \left[\sum_{i=1}^N \sum_{j=1}^N \mathbf{X}_{n-i}^T \mathbf{A}_i^T \mathbf{A}_j \mathbf{X}_{n-j} \right] \end{aligned} \quad (4.29)$$

Making use of the trace property for a $k \times k$ matrix $\mathbf{B}$ and two k -dimensional vectors $\mathbf{u}$ and $\mathbf{v}$ such that $\mathbf{u}^T \mathbf{B} \mathbf{v} = \text{Tr}(\mathbf{B} \mathbf{v} \mathbf{u}^T)$, (4.29) can be rewritten as:

$$D_{\text{MSE}} = R_{00} + 2\text{Tr}(\mathbf{A}^T \mathbf{V}) + \text{Tr}(\mathbf{A}^T \mathbf{R} \mathbf{A}) \quad (4.30)$$

where $R_{00} = E[\|\mathbf{X}_n\|^2] = E[\mathbf{X}_n^T \mathbf{X}_n]$. For optimal prediction, the Minimum MSE (MMSE) can be found by substituting (4.28) into (4.30):

$$D_{\text{MSE}(\text{min})} = R_{00} + \text{Tr} \left(\sum_{j=1}^N \mathbf{A}_j \mathbf{R}_{j0} \right) \quad (4.31)$$

Assuming ergodicity holds for the vector process $\{\mathbf{X}_n\}$, the expectation operator $E[\cdot]$ can be approximated by time averaging a sufficiently large number of observed vectors, say $M \gg N$. In this case, the LP coefficient matrices: $\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N$ are chosen to minimize the prediction residual energy over M and (4.21) can be approximated by:

$$D_{\text{MSE}} = \frac{1}{M} \sum_{n=1}^M \|\mathbf{e}_n\|^2 = \frac{1}{M} \sum_{n=1}^M \mathbf{e}_n^T \mathbf{e}_n \quad (4.32)$$

The above approximation thus leads to two approaches in estimating the supermatrix element $\mathbf{R}_{ij}$. In the vector covariance method, $\mathbf{R}_{ij}$ is represented by $\tilde{\mathbf{R}}_{ij}$ defined as a series

sum of vector cross-products:

$$\tilde{\mathbf{R}}_{ij} = \sum_{n=N}^M \mathbf{X}_{n-i} \mathbf{X}_{n-j}^T \quad (4.33)$$

Note that the time index n starts at N to avoid using vector values outside the observation interval $[1, 2, \dots, M]$. Thus making use of (4.33), (4.27) can be approximated by:

$$\tilde{\mathbf{R}}\mathbf{A} = -\tilde{\mathbf{V}} \quad (4.34)$$

where $\tilde{\mathbf{R}}$, and $\tilde{\mathbf{V}}$ are approximations to the supermatrix $\mathbf{R}$ and supervector $\mathbf{V}$ and $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N]^T$ is the LP supervector. In a similar way, the scalar covariance method is easily obtained from (4.33) by setting the vector dimension k in (4.16) to 1.

In the vector autocorrelation method, $\tilde{\mathbf{R}}_{ij}$ is considered as an infinite series sum of vector cross-products but with the assumption that $\mathbf{X}_n = \mathbf{0}$ if n is outside the observation interval. Thus $\tilde{\mathbf{R}}_{ij}$ can be expressed as:

$$\tilde{\mathbf{R}}_{ij} = \tilde{\mathbf{R}}_{|i-j|} = \sum_{n=1+|i-j|}^M \mathbf{X}_n \mathbf{X}_{n-|i-j|}^T \quad (4.35)$$

The $\tilde{\mathbf{R}}_{ij}$ matrices are not necessarily Toeplitz except when $i=j$. In other words, the supermatrix $\tilde{\mathbf{R}}$ can be only described as block-Toeplitz and symmetric. Special numerical methods are available to solve for the set of predictive coefficient matrices and will be discussed in the next section.

A special variation of VLP which is of use in the development of the parametric extrapolation and interpolation techniques is the affine VLP [19]. An underlying assumption in the previous VLP derivation is that the stochastic vectors always have zero mean. This is not necessarily true if VLP is applied to some non-negative processes. Normal form of VLP can be applied but would yield inferior performance. Instead the following affine VLP can be employed:

$$\mathbf{Y} = -\sum_{j=1}^N \mathbf{A}_j \mathbf{X}_j + \mathbf{C} \quad (4.36)$$

where $\mathbf{C}$ is a constant column vector defined as:

$$C = E[Y] + \sum_{j=1}^N A_j E[X] \quad (4.37)$$

By substituting C into (4.31), a more recognizable form can be obtained as:

$$Y - E[Y] = - \sum_{j=1}^N A_j (X_j - E[X]) \quad (4.38)$$

In the prediction process, the expected value $E[X]$ is subtracted from the observed vector X_j and the expected value $E[Y]$ is appended after the summation. For most purposes, the predicted vector Y and the observed vectors X_j belong to the same stochastic vector set and their expected mean values can be best approximated by:

$$E[Y] = E[X] \approx \frac{1}{M} \sum_{j=1}^M X_j \quad (4.39)$$

4.2.1.1 Computational Issues

Solution to the above vector linear predictor can be treated the same as solving a special linear system which has the same number of equations as well as unknowns such that the system is neither overdetermined nor underdetermined. There are two issues which need to be addressed in the VLP formulation. First of all, the linear system can be moderately large because the number of supermatrix elements is given by $N \cdot k \times N \cdot k$ (noting N is the VLP order and k is the vector dimension). Moreover, by deriving (4.34) directly in terms of the data matrix Z :

$$Z^T Z A = -Z^T Y \quad (4.40)$$

where each row of Z consists of N observed vectors X_j :

$$Z = \begin{bmatrix} X_0^T & X_{-1}^T & \cdots & X_{(1-N)}^T \\ X_1^T & X_0^T & \cdots & X_{(2-N)}^T \\ \vdots & \vdots & \cdots & \vdots \\ X_{(M-1)}^T & X_{(M-2)}^T & \cdots & X_{(M-N)}^T \end{bmatrix} \quad (4.41)$$

2 of/de 2

MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

and $\mathbf{Y}$ is a supervector denoting the M observed vectors:

$$\mathbf{Y} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_M]^T \quad (4.42)$$

the system could possibly be numerically ill-conditioned [9] because of round-off errors in $\mathbf{A}$ and $\tilde{\mathbf{V}}$.

An effective approach to alleviate the ill-conditioning problem in high order LPC analysis is by adding "noise" to the system. The idea is to add a constant to the major diagonal matrix elements such that $\tilde{\mathbf{R}} = \tilde{\mathbf{R}} + \sigma \mathbf{I}$. The constant (noise energy) σ is usually chosen as a scale factor of the mean value of the diagonal elements (mean signal energy). This has the effect of bandlimiting the input signal range and moving the poles of LPC filter towards the origin. Additionally, a number of numerically stable algorithms can be used in obtaining the vector covariance and vector autocorrelation solutions. For the covariance method, the decomposition type algorithms [53] which include Cholesky decomposition, QR decomposition (via the Householder transform) and Singular Value Decomposition (SVD) may be used. Instead of performing direct inversion of the supermatrix $\mathbf{A}$, these algorithms break up the original matrix into more elementary form for easier manipulation. The Cholesky method which is particularly well-suited for handling symmetric matrices is chosen in this study (see Appendix A for the VLP solution via Cholesky decomposition). In the autocorrelation method where the matrix has block-Toeplitz property, a generalized version of the Levinson-Durbin algorithm called Levinson-Wiggins-Robinson (LWR) algorithm [63][4] can be employed. The LWR method is described in Appendix B.

4.2.2 CELP Frame Reconstruction

In this section, the suitability and results of applying the VLP method for the reconstruction of ICELP speech parameters are discussed. Though not of direct relevance to this research study, it is noted the VLP technique has been applied successfully to inter-frame LPC coding [64] and Vector Adaptive Predictive Coding (VAPC) [12].

Use of parametric interpolation is not new in the ICELP design. To reduce the bit rate, the frame gain and LPC parameters are coded only once per frame. Since an ICELP frame is partitioned into four subframes to accommodate rapid signal variation over the frame

interval the LPC values at the subframe boundary are linearly interpolated using values from previous and current frames to improve performance. This approach can be extended to lost frame reconstruction. The lost frame parameters are interpolated from the previous and future parameters via VLP rather than the sub-optimal linear method. Likewise in the extrapolation, the lost parameters are estimated solely using parameters from the previous frames.

To assist the design of the reconstruction algorithm, a perceptual ranking was obtained for the ICELP parameters by randomly substituting the value of a particular ICELP parameter with its last frame value while keeping the other parameters intact. The decoded speech can then be listened to determine the relative impact on speech quality. Similar to the VSELTP and RPE-LTP, the frame gain is perceptually the most important parameter, followed by the short-term LPC filter information, ACB lags, ACB coefficients, stochastic excitation gain values and stochastic codebook indices. The reconstruction of these ICELP parameters are next discussed in the order of decreasing perceptual importance.

4.2.2.1 Frame Gain

The ICELP frame gain, f_n , is a product of the residual energy of the MSVQ short-term LPC filter and the speech energy and is coded once per frame. An estimate of f_n in the n^{th} lost frame, $\hat{f}_n$, can be obtained from either one of the following:

$$\begin{aligned}\hat{f}_n &= a_{1s,1}^{(f)} f_{n-1} + a_{1s,2}^{(f)} f_{n-2} \\ \hat{f}_n &= a_{2s,1}^{(f)} f_{n+1} + a_{2s,2}^{(f)} f_{n-1}\end{aligned}\tag{4.43a,b}$$

where $\{a_{1s,1}^{(f)}, a_{1s,2}^{(f)}\}$ are the 1-sided LP scalar coefficients of the extrapolation function and $\{a_{2s,1}^{(f)}, a_{2s,2}^{(f)}\}$ are the 2-sided LP coefficients of the interpolation function. Table 4.1 shows the values obtained for the above LP coefficients.

$a_{1s,1}^{(f)}$	$a_{1s,2}^{(f)}$	$a_{2s,1}^{(f)}$	$a_{2s,2}^{(f)}$
0.8842	-0.1260	0.5010	0.5010

Table 4.1 - Frame Gain LP Coefficients

Solutions to these two sets of LP coefficients proceed by formulating the Wiener-Hopf

equations via covariance analysis on a time series of 25,000 quantized frame gain values from the TIMIT database. The formulation of the extrapolator and interpolator differ in the way the frame gain time series is accessed. Whereas the extrapolator uses consecutive frame gain values as its “training sequence”, the interpolator uses only the alternate values. A softening factor which is equivalent to $(0.01 \times \text{mean value of all matrix elements } \tilde{\mathbf{R}}_{ii})$ is added to the major diagonal of the covariance matrix and the equations are solved by Cholesky decomposition. Since f_n is always non-negative, affine LP is used such that a long-term frame gain average $\tilde{f}$ ($=0.0921$) is removed from all frame gain values during the covariance analysis and added back during extrapolation and interpolation. Any negative frame gain estimate is always replaced by the smallest quantized frame gain value.

Figure 4.5 shows a time series plot of actual and predicted frame gain values based on extrapolation using a 2nd order affine LP. Except on several occasions where the LP estimates differ from the original frame gain by a large margin, the extrapolator can track the original value given last two frame gain values are correctly received. It can also be observed from the graph that for the small frame gain values, the predicted values always float above the actual values and vice versa for large frame gain values. This can be explained by the fact that the linear predictor always try to place the LP estimates as “close” as possible to $\tilde{f}$ for MSE minimization.

4.2.2.2 Short-term LPC Filter

Coding of short-term LPC parameters is crucial to the performance of a low-bit rate speech coder [10]. Accurate reconstruction of the LPC parameters (at least in the perceptual sense) is necessary to minimize output speech distortion. Since successive LPC parameters have significant correlation because of the slow variation of the short-term speech spectrum, it is feasible to reconstruct the lost LPC parameters based on the previous and possibly future frames.

In (2.1), the short-term LPC filter is defined in terms of a set of p time-varying filter coefficients $a_{n,k}$. The LPC coefficients are rather sensitive to quantization error and are usually transformed into another equivalent form such as reflection coefficients (otherwise known as the partial correlation coefficients) or Line Spectrum Pairs for speech coding purposes. Use of these parameters for voice frame reconstruction is now examined.

Figure 4.5 - Second Order LP-based Extrapolation on Frame Gain

The reflection coefficients which have a physical interpretation as the parameters of an acoustic tube model of speech can be obtained as a by-product of the autocorrelation solution to the Wiener-Hopf equations [19]. There is a unique mapping between the LPC coefficients $\{a_k\}$ and reflection coefficients $\{k_i\}$. A step-up procedure can be used to convert from reflection coefficients k_i to LPC coefficients (of a p^{th} order short-term filter) $a_i^{(p)}$ as shown below:

$$\begin{aligned} a_i^{(i)} &= k_i, \\ a_j^{(i)} &= a_j^{(i-1)} + k_i a_{i-j}^{(i-1)}, \quad \begin{matrix} i = 1, \dots, p \\ j = 1, \dots, i-1 \end{matrix} \end{aligned} \quad (4.44)$$

while a step-down procedure can be used to convert the two forms in the opposite way:

$$\begin{aligned} k_i &= a_i^{(i)}, \\ a_j^{(i-1)} &= \frac{a_j^{(i)} - a_i^{(i)} a_{i-j}^{(i)}}{1 - k_i^2}, \quad \begin{matrix} i = p, \dots, 1 \\ j = 1, \dots, i-1 \end{matrix} \end{aligned} \quad (4.45)$$

For a stable LPC filter (with all its poles inside the unit circle), the reflection coefficients are bounded such that:

$$-1 < k_i < 1, \quad i = 1, \dots, p \quad (4.46)$$

The unit magnitude bound makes it easier to quantize the reflection coefficients and at the same time (4.47) provides a stability test for the short-term LPC filter. Rather than converting to direct form filter coefficients, the reflection coefficients can be used to derive a lattice filter structure for direct speech synthesis as is done in the VSELP and RPE-LTP speech decoders [14][16].

Another representation of LPC parameters which have gained much interest in recent years is the Line Spectrum Pairs (LSPs) or Line Spectral Frequencies (LSFs) [44]. Associated with the LSP formulation are two polynomials $P(z)$ and $Q(z)$ where

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)} A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)} A(z^{-1}) \end{aligned} \quad (4.47a,b)$$

where $A(z)$ is the inverse LPC filter as defined in (2.1). The following properties are noted

from this pair of polynomials:

1. $P(z)$ and $Q(z)$ are symmetric and antisymmetric polynomials and if the respective polynomial coefficients are denoted by p_i and q_i , then

$$p_i = p_{p+1-i}, \quad q_i = -q_{p+1-i}$$

2. All zeros of $P(z)$ and $Q(z)$ lie on the unit circle and can be expressed as $e^{j\omega_i}$, where ω_i , $1 \leq i \leq p$, are the LSFs.
3. For each polynomial, there are p zeros which form $p/2$ complex conjugate pairs. The $2p$ zeros of $P(z)$ and $Q(z)$ can be uniquely represented using p LSFs, $\omega_1, \omega_2, \dots$ and ω_p , in the upper half of the unit circle.
4. The zeros of $P(z)$ and $Q(z)$ alternate on the unit circle. This is a necessary and sufficient condition for the stability of $A_n(z)$ and should be observed in the quantization of the zeros.

Roots finding can take advantage of the unit circle property and the interleaving property of zeros. An efficient method to compute the zeros employing Chebyshev polynomials is described in [30].

In utilizing the VLP, the decoded reflection coefficients and LSPs are treated as vectors $\mathbf{k}_n$ and ω_n such that:

$$\begin{aligned} \mathbf{k}_n &= [k_{1,n}, k_{2,n}, \dots, k_{p,n}] \\ \omega_n &= [\omega_{1,n}, \omega_{2,n}, \dots, \omega_{p,n}] \end{aligned} \quad (4.48a,b)$$

So far the VLP assumes the use of MSE criteria as a distortion measure. Based on the fact that the short-term LPC filter can be treated as a model for the speech spectral envelope, another VLP distortion measure known as the Likelihood Ratio (LR) [35][39] can be formulated for the LPC coefficients and herein explained.

If $D(z)$ is the Z transform of the residual excitation d_n , then Parseval theorem states that the total residual energy, α , can be related via the following:

$$\alpha = \sum_{n=-\infty}^{\infty} d_n^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |D(e^{j\omega})|^2 d\omega \quad (4.49)$$

As d_n is obtained by inverse filtering the speech input s_n such that $D(z) = S(z)A(z)$, it is possible to express (4.50) as:

$$\alpha = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S(e^{j\omega})|^2 |A(e^{j\omega})|^2 d\omega \quad (4.50)$$

For a p^{th} order linear predictor, (4.47) can be rearranged in a more convenient form via the following:

$$\begin{aligned} \alpha &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |S|^2 |A|^2 d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |S|^2 \left[1 + \sum_{k=1}^p a_k e^{j\omega k} \right] \left[1 + \sum_{i=1}^p a_i e^{-j\omega i} \right] d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |S|^2 d\omega + \frac{1}{2\pi} \sum_{i=1}^p a_i \int_{-\pi}^{\pi} |S|^2 e^{j\omega i} d\omega \\ &\quad + \frac{1}{2\pi} \sum_{k=1}^p a_k \int_{-\pi}^{\pi} |S|^2 e^{j\omega k} d\omega + \frac{1}{2\pi} \sum_{i=1}^p \sum_{k=1}^p a_i a_k \int_{-\pi}^{\pi} |S|^2 e^{-j\omega(i-k)} d\omega \\ &= r(0) + 2 \sum_{k=1}^p a_k r(k) + \sum_{k=1}^p \sum_{i=1}^p a_k a_i r(i-k) \end{aligned} \quad (4.51)$$

where $S = S(e^{j\omega})$ and $A = A(e^{j\omega})$. The above expression can be recognized to be the MSE expression for LPC filter (see (4.30)). It can be expressed in matrix form as:

$$\alpha = r(0) + 2\mathbf{a}^T \mathbf{r} + \mathbf{a}^T \mathbf{R}_p \mathbf{a} = \mathbf{a}'^T \mathbf{R}_p \mathbf{a}' \quad (4.52)$$

where $\mathbf{a}' = [1 \ a_1 \ a_2 \ \dots \ a_p]$ is the augmented LPC coefficient vector and $\mathbf{R}_p$ is the Toeplitz autocorrelation matrix with first row $\mathbf{r}_p = [r(0) \ r(1) \ \dots \ r(p)]$. For the optimal p^{th} order LP, the minimum residual energy α_p is the scalar equivalent of (4.31) which is given by:

$$\alpha_p = \mathbf{a}_p'^T \mathbf{r}_p$$

If the optimal p^{th} order inverse short-term filter $A_p(z)$ has an augmented coefficient vector $\mathbf{a}'_p$, then the Lk between a particular p -dimensional LPC coefficient vector $\mathbf{a}'$ and

$\mathbf{a}'_p$ is defined as one less the ratio of the respective residual energies:

$$\begin{aligned} d_{LR}(\mathbf{a}', \mathbf{a}'_p) &= \frac{\alpha}{\alpha_p} - 1 \\ &= \frac{\mathbf{a}'^T \mathbf{R}_p \mathbf{a}'}{\mathbf{a}'_p^T \mathbf{R}_p \mathbf{a}'_p} - 1 = \frac{\mathbf{a}'^T \mathbf{R}_p \mathbf{a}'}{\mathbf{a}'^T \mathbf{r}_p} - 1 \end{aligned} \quad (4.53)$$

The $(p+1)$ autocorrelation coefficients of the row vector $\mathbf{r}_p$ can be obtained from the impulse response of the synthesis filter using the correlation matching method described in [40]. The LR has the usual properties of a distortion measure. It is always non-negative $d_{LR}(\mathbf{a}', \mathbf{a}'_p) \geq 0$ and the minimum distortion is zero when $\mathbf{a}' = \mathbf{a}'_p$. However, it should be noted LR is neither a distance nor a metric. In Section 5.2.3, it will be shown the LR is an approximation to the log-spectral distance measure.

A VLP scheme can be determined based on minimizing the average LR given by:

$$\begin{aligned} \bar{d}_{LR} &= \min \left[\frac{1}{M} \sum_{n=1}^M d_{LR}(\hat{\mathbf{a}}'_n, \mathbf{a}'_{p,n}) \right] \\ &= \min \left[\frac{1}{M} \sum_{n=1}^M \left(\frac{r(0) + 2\hat{\mathbf{a}}_n^T \mathbf{r} + \hat{\mathbf{a}}_n^T \mathbf{R}_p \hat{\mathbf{a}}_n}{\alpha_{p,n}} - 1 \right) \right] \end{aligned} \quad (4.54)$$

where $\hat{\mathbf{a}}'_n$ and $\hat{\mathbf{a}}_n$ are the augmented and unaugmented versions of the reconstructed LPC coefficient vectors for the n^{th} lost frame. For 2nd order VLP, $\hat{\mathbf{a}}_n$ can be estimated by either one of following:

$$\begin{aligned} \hat{\mathbf{a}}_n &= \mathbf{M}_{1s,1} \mathbf{a}_{n-1} + \mathbf{M}_{1s,2} \mathbf{a}_{n-2} \\ \hat{\mathbf{a}}_n &= \mathbf{M}_{2s,1} \mathbf{a}_{n+1} + \mathbf{M}_{2s,2} \mathbf{a}_{n-1} \end{aligned} \quad (4.55a,b)$$

where $\{\mathbf{M}_{1s,1}, \mathbf{M}_{1s,2}\}$ and $\{\mathbf{M}_{2s,1}, \mathbf{M}_{2s,2}\}$ are the respective sets of $p \times p$ square LP matrices. Consider the 1-sided VLP scheme in (4.55a), it is possible to express it in the following manner:

$$\begin{bmatrix} \hat{a}_{1,n} \\ \hat{a}_{2,n} \\ \vdots \\ \hat{a}_{p,n} \end{bmatrix} = \begin{bmatrix} m_{11}^{(1)} & m_{12}^{(1)} & \dots & m_{1p}^{(1)} & m_{11}^{(2)} & m_{12}^{(2)} & \dots & m_{1p}^{(2)} \\ m_{21}^{(1)} & \dots & \dots & \dots & m_{21}^{(2)} & \dots & \dots & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ m_{p1}^{(1)} & \dots & \dots & m_{pp}^{(1)} & m_{p1}^{(2)} & \dots & \dots & m_{pp}^{(2)} \end{bmatrix} \begin{bmatrix} a_{1,n-1} \\ a_{2,n-1} \\ \vdots \\ a_{p,n-1} \\ a_{1,n-2} \\ a_{2,n-2} \\ \vdots \\ a_{p,n-2} \end{bmatrix} \quad (4.56)$$

where $m_{ij}^{(k)}$ and $a_{i,k}$ denote the LP matrix element of $\mathbf{M}_{1s,k}$ and element of LPC vector $\mathbf{a}_k$. If the matrix and vector are interchanged on the right hand side of (4.56), then the equation can be rewritten compactly in matrix form as:

$$\hat{\mathbf{a}}_n = \Lambda_{n-1,n-2} \mathbf{m} \quad (4.57)$$

where $\Lambda_{n-1,n-2}$ is the $2p^2 \times 2p^2$ matrix with elements from the vectors $\mathbf{a}_{n-1}$ and $\mathbf{a}_{n-2}$ and $\mathbf{m}$ is the likelihood LP vector consisting of elements from the matrices $\mathbf{M}_{1s,1}$ and $\mathbf{M}_{1s,2}$.

The VLP minimization thus involves finding the best values of $\hat{\mathbf{a}}_n$ in the numerator expression in Equation (4.55a). If $\hat{\mathbf{a}}_n$ is replaced by Equation (4.57) in the numerator, then

$$\begin{aligned} \bar{d}_{LR} &= r_0 + 2\hat{\mathbf{a}}_n^T \mathbf{r} + \hat{\mathbf{a}}_n^T \mathbf{R} \hat{\mathbf{a}}_n \\ &= r_0 + 2\mathbf{m}^T \Lambda_{n-1,n-2}^T \mathbf{r} + \mathbf{m}^T \Lambda_{n-1,n-2}^T \mathbf{R} \Lambda_{n-1,n-2} \mathbf{m} \end{aligned} \quad (4.58)$$

If we further define a new transformed vector as $\mathbf{u}_{n-1,n-2} = \Lambda_{n-1,n-2}^T \mathbf{r}$, the above then becomes:

$$\begin{aligned} \bar{d}_{LR} &= r_0 + 2\mathbf{m}^T (\Lambda_{n-1,n-2}^T \mathbf{r}) + \mathbf{m}^T (\Lambda_{n-1,n-2}^T \mathbf{r} \mathbf{r}^T \Lambda_{n-1,n-2}) \mathbf{m} \\ &= r_0 + 2\mathbf{m}^T \mathbf{u}_{n-1,n-2} + \mathbf{m}^T (\Lambda_{n-1,n-2}^T \mathbf{r}) (\Lambda_{n-1,n-2}^T \mathbf{r})^T \mathbf{m} \\ &= r_0 + 2\mathbf{m}^T \mathbf{u}_{n-1,n-2} + \mathbf{m}^T \mathbf{U}_{n-1,n-2} \mathbf{m} \end{aligned} \quad (4.59)$$

where $\mathbf{U}_{n-1, n-2} = \mathbf{u}_{n-1, n-2} \mathbf{u}_{n-1, n-2}^T$. Since (4.59) has the same quadratic form as (4.29) and $\mathbf{U}_{n-1, n-2}$ is a symmetric matrix, the above expression can be solved via Cholesky decomposition. A noise softening factor which is computed in the same manner as described in Section 4.2.2.1 is added to alleviate possible ill-conditioning problem. The solved likelihood LP vector ($\mathbf{m}$) can then be used to extrapolate the lost LPC coefficients via (4.56).

LPC Parameters	VLP	Minimization Criteria
K = Reflection Coefficient	1S = 1-sided VLP Extrapolation	LR = Likelihood Ratio
LSP= Line Spectrum Pairs	2S = 2-sided VLP Interpolation	MSE = Mean Squared Error

Table 4.2 - VLP Reconstruction Methods for LPC Parameters

Table 4.2 summarizes the notations adopted for various VLP methods employed in this study. For example, LR-LSP-VLP-2S refers to a 2-sided VLP scheme based on minimizing the LR criteria whereby the LSPs are always converted to LPC coefficients prior to VLP. Figure 4.6 and Fig. 4.7 show typical time series plots of short-term LPC parameters based on MSE-K-VLP-1S and LR-LSP-VLP-1S. In Fig. 4.6, the n^{th} reflection coefficient vector is extrapolated based on the $(n-1)^{\text{st}}$ and $(n-2)^{\text{nd}}$ vectors using a MSE criteria. In Fig. 4.7, the n^{th} LPC coefficient vector is extrapolated based on the last two LPC coefficient vectors which are obtained by transforming the LSPs to LPC coefficients. Although both VLP methods appear to provide good estimates of the LPC parameters as shown in the diagrams, there is significant difference in the reconstructed speech quality between the two schemes. The objective and subjective performance of the various VLP reconstruction methods is discussed in details in Chapter 5.

Unlike the autocorrelation method for VLP, the covariance method does not necessarily guarantee the stability of the short-term LPC filter; the poles of $1/A(z)$ could be outside the unit circle. In MSE-K-VLP, stability could be guaranteed by ensuring the reconstructed reflection coefficients are within the bounds as shown in (4.47). If this is not the case, the reflection coefficients are clipped. At the same time, the reconstructed LSFs are reordered to ensure stability for MSE-LSP-VLP. For LR-K-VLP and LR-LSP-VLP, stability can be ensured in a two-step manner. First, the reconstructed LPC coefficients are

Figure 4.6 - Second Order MSE-K-VLP-IS Extrapolation

Figure 4.7 - Second Order LR-LSP-VLP-1S Extrapolation

de-emphasized using a bandwidth expansion factor $\gamma \approx 1 - \epsilon$ (where ϵ is a small value) such that the new inverse LPC filter estimate becomes:

$$\hat{A}'(z) = \hat{A}(z/\gamma) \quad (4.60)$$

where $\hat{A}(z)$ is the original filter estimate. The expansion factor has the effect of moving the poles towards the origin and is equivalent to spreading the formant regions in the speech spectrum. After repeated experimental trials, γ is chosen to be 0.985. Following the de-emphasis, the LPC coefficients are converted into the reflection coefficient domain using the step-down procedure in (4.45) and any unstable reflection coefficients are reduced to ± 0.99 . The stabilized values can be converted back to the LPC coefficients for speech synthesis afterwards.

4.2.2.3 Adaptive Codebook

The adaptive codebook parameters include a lag or ACB index (m) and a 3-dimensional ACB coefficient vector ($\mathbf{b}$). Time series of typical lag and ACB coefficients and corresponding MSE-VLP-1S estimates are shown in Fig. 4.8 and Fig. 4.9. Except for b_0 , the extrapolator performed poorly for the lag m and the other two ACB coefficients $\{b_1, b_2\}$ as they all exhibit erratic behavior in their time series. In the case of lag, the ICELP encoder can sometimes lock onto multiples of the fundamental pitch period during ACB search and results in pitch doubling or tripling effects [10]. It is also noted that the large lags often correspond to an unvoiced speech frame.

It was deduced that lag smoothness is important to minimize distortion in the reconstructed frame during frame losses. In this study, a modified pitch search strategy similar to the U.S. FS-1016 CELP coder is employed in the ICELP design [2][3]. Referring to the closed-loop WMSE search criteria in (2.15), it is possible to define a normalized WMSE criteria $\epsilon_{\text{norm}}^{(m)}$:

$$\epsilon_{\text{norm}}^{(m)} = \frac{\epsilon^{(m)}}{\mathbf{y}_1^T \mathbf{y}_1} = 1 - \frac{|\mathbf{c}_1 \mathbf{D}^{(m)}|^2}{|\mathbf{H} \mathbf{D}^{(m)}|^2 |\mathbf{y}_1^T \mathbf{y}_1|} \quad (4.61)$$

An initial lag value m is obtained by maximizing the right most fractional term in (4.61). The WMSE is normalized and compared to an unvoiced threshold of -3dB. If the

Figure 4.8 - Comparison of Actual and Predicted Lag Values

Figure 4.9 - Comparison of Actual and Predicted ACB Coefficients

normalized value is above the threshold, the modified search strategy then examines the normalized WMSEs in the vicinity (± 10) of $m/2$, $m/3$ and $m/4$ whose WMSE values were computed during the previous exhaustive ACB search. The smallest lag value which has WMSE within 1dB from the initial WMSE is chosen for final decoding instead. The initial lag value is retained if either the WMSE is below the threshold or sub-multiple lags all have much greater WMSE. Figure 4.10 shows a comparison of the lag contour with and without the "modified" search strategy. A 0.5-1.0 dB decrease in average segmental SNR was noted for the encoded speech with a smoothed lag contour.

At the decoder, a non-linear median filter [48] can be further applied to assist in the lag reconstruction. For extrapolation of the n^{th} frame, a 3-point median filter is applied to the last three lag values of the $(n-1)^{\text{st}}$ frame and the median is used all through the n^{th} frame. For interpolation of the n^{th} frame, a 7-point median is used with the last 4 lags from the $(n-1)^{\text{st}}$ frame and first 3 lags from the $(n+1)^{\text{st}}$ frame.

Although these two methods have significantly improved reconstruction of the lag values, reconstruction of the ACB coefficients remains an issue. Since the VLP method is less effective in estimating the lost ACB coefficients, an approach based on conditional repeat of parameters is adopted instead. To ensure that continuous repetition of the ACB coefficients do not lead to unstabilized filter response, the initial pitch gain codevectors are checked for stability using a procedure described in [50]. Unstable ACB coefficient vectors are modified accordingly to minimize loss in prediction gain.

The reconstruction method makes a decision if the lost frame is voiced. This can be done by comparing the median to the four lags (in VLP-1S) or seven lags (in VLP-2S) and if over half of the lags are within ± 5 samples from the median value, then the lost frame is estimated to be voiced. The ACB coefficient vector can be reconstructed via one of the following:

1. If the n^{th} frame is estimated to be unvoiced, the reconstruction scheme then selects a scaled version of coefficient vector associated with the lag median.
2. If the n^{th} frame is voiced and the extrapolation method is employed, a scaled version of the coefficient vector of the last subframe of the $(n-1)^{\text{st}}$ frame is used.
3. If the n^{th} frame is voiced and the interpolation method is employed, a scaled

Figure 4.10 - Lag Contour Comparison using Modified Pitch Search

version of the coefficient vector from either the last subframe of the $(n-1)^{\text{st}}$ frame or the first subframe of the $(n+1)^{\text{st}}$ frame could be used depending upon whether the lag median comes from the $(n-1)^{\text{st}}$ frame or the $(n+1)^{\text{st}}$ frame.

A scale factor of 0.90 is used in all of the above. This intuitive method thus takes into account the ACB coefficients in the immediate subframes (on either side of the lost frame) and provides a good estimate of the lost values.

4.2.2.4 Stochastic Excitation Gain

Like the ACB parameters, the stochastic codebook gain values are generated 4 times per frame and are treated as a 4-dimensional vector $\mathbf{g}_n = [g_{1n}, g_{2n}, g_{3n}, g_{4n}]$. VLP is applied to obtain the vector estimate $\hat{\mathbf{g}}_n$ in the lost frame via one of the following:

$$\begin{aligned}\hat{\mathbf{g}}_n &= \mathbf{A}_{1s,1}^{(c)} \mathbf{g}_{n-1} + \mathbf{A}_{1s,1}^{(g)} \mathbf{g}_{n-2} \\ \hat{\mathbf{g}}_n &= \mathbf{A}_{2s,1}^{(g)} \mathbf{g}_{n+1} + \mathbf{A}_{2s,1}^{(g)} \mathbf{g}_{n-1}\end{aligned}\tag{4.63a,b}$$

where $\{\mathbf{A}_{1s,1}^{(g)}, \mathbf{A}_{1s,2}^{(g)}\}$ and $\{\mathbf{A}_{2s,1}^{(g)}, \mathbf{A}_{2s,2}^{(g)}\}$ are the respective sets of 4×4 LP coefficient matrices of the extrapolator and interpolator. Similar to frame gain prediction, 2^{nd} order affine VLPs were obtained via covariance analysis on a time series of 25,000 quantized stochastic gain vectors. A long-term mean vector $\bar{\mathbf{g}} = \{0.0390, 0.0390, 0.0390, 0.0390\}$ is subtracted from every input vector prior to covariance analysis and added back during the extrapolation and interpolation. After formulating the Wiener-Hopf equations in (4.34), the major diagonal of the supermatrix is appended with a softening factor equal to $(0.01 \times \text{mean value of all matrix elements } \tilde{\mathbf{R}}_{ii})$ and then the equations are solved via the vector version of Cholesky decomposition in Appendix B. To ensure the VLP always lead to positive stochastic gain values, any negative stochastic gain estimate is always replaced by the smallest entry in the quantization table.

Figure 4.11 shows a time series plot of actual and extrapolated stochastic gain values based on 2^{nd} order affine VLP. It can be seen that although the predicted vector $\hat{\mathbf{g}}_n$ cannot reproduce the finer details of the stochastic gain variation (and in particular the small gain values), it can track the gradual changes in the stochastic gain contour and is deemed adequate for the reconstruction purposes.

Figure 4.11 - Stochastic Gain Extrapolation via Second Order VLP

4.2.2.5 Stochastic Codebook

The ICELP stochastic codebook parameters consist of two sets of codebook indices and sign multipliers. Based on the fact the stochastic codebook vectors are usually modelled by white Gaussian noise in CELP coder design, the time series of either parameters have unpredictable behavior because of relatively little correlation in between frames. Neither extrapolation nor interpolation is appropriate for reconstructing these parameters.

The initial parameter classification indicate that errors in the stochastic codebook parameters only result in noisier sounding speech and are not perceptually disturbing, a fact further confirmed in [10][3]. Consequently, the stochastic codebook parameters are chosen randomly for the lost frames.

4.2.2.6 Adaptive Energy Control

Speech energy control is a very important part of the extrapolation and interpolation methods. In case of gross reconstruction errors which may arise in the short-term LPC and gain parameters, the mechanism tends to suppress perceptually disturbing squeaks and loud and clipped speech.

As noted in Section 4.2.2, the ICELP design incorporates intraframe interpolation of the LPC and frame gain parameters. Hence the reconstructed errors are already smoothed to some degree. Furthermore, the adaptive energy control always ensure the synthesized speech energy does not exceed the previous correctly received frames. From (2.7), the ICELP excitation is given by:

$$\begin{aligned}\hat{\mathbf{d}}^{(m)} &= \mathbf{G}^{(l)} \mathbf{v}^{(l)} + \mathbf{D}^{(m)} \mathbf{b}^{(m)} \\ &= (f_n \cdot g_n) \mathbf{v}^{(l)} + \mathbf{D}^{(m)} \mathbf{b}^{(m)}\end{aligned}\tag{4.64}$$

where $\mathbf{v}^{(l)}$ is the stochastic codebook excitation and the term $\mathbf{D}^{(m)} \mathbf{b}^{(m)}$ is the ACB contribution as defined in (2.17). The synthetic speech vector $\hat{\mathbf{s}}_n$ of the n^{th} subframe is generated by convolving the residual excitation $\hat{\mathbf{d}}^{(m)}$ with $1/A(z)$. If the subframe length is K_c , the total speech energy (of the n^{th} subframe) can be computed as:

$$E_n = \sum_{k=0}^{K_c-1} \hat{s}_{n,k}^2\tag{4.65}$$

The adaptive energy control thus compares current subframe energy with that in the previous subframe and if $E_n \geq E_{n-1}$, the synthesized speech is “undone”. As a result, the frame gain f_n and adaptive codebook contribution $D^{(m)} b^{(m)}$ are scaled by $\alpha \sqrt{E_{n-1}/E_n}$ and $\beta \sqrt{E_{n-1}/E_n}$ respectively (where $\alpha = 0.975$ and $\beta = 0.950$) and the decoded speech is resynthesized. No sudden energy explosion can therefore result. Figure 4.12 shows the use of adaptive energy control to correct a reconstructed speech segment which has an annoying pop.

4.2.2.7 Multiple Lost Frames

Although the extrapolation and interpolation methods are based on the assumption the lost frame usually occurs in an isolated manner, more than one frame can be lost in practice. Since VLP reconstruction is based on the short-term stationarity of speech, VLP becomes ineffective when there are three or more frame losses. On such occasion, the frame gain and LPC parameters from the second reconstructed frame are attenuated in 4dB decrements and bandwidth expanded as per (2.3) until the speech is completely muted.

4.3 Summary

In this chapter, a number of PCM speech reconstruction schemes were considered and found to be adequate even at high loss rates. However, it was shown that low-bit rate speech coders in general can tolerate a much lower frame error rate because of the memory effect required for adaptation. Lost frame recovery methods based on last frame repetition are often used in CELP coders. Alternative frame reconstruction methods using VLP-based extrapolation and interpolation were described. Objective and subjective results and comparison to the last-frame repetition method will be given in the next chapter.

Figure 4.12 - Adaptive Energy Control Mechanism

Chapter 5

Measurements of Speech Quality

5.1 Introduction

One of the most difficult tasks in speech coding is the assessment of relative performance of different coding algorithms. The major problem lies in the fact it has not been possible to objectively quantify the term “speech quality”. While some listeners may tolerate one kind of distortion, others may not. Further complication arises as the human listener has a built-in adaptation mechanism. Noisy speech which was unsatisfactory in the beginning can become acceptable after repeated listening.

Although no single quality measurement method exists, a combination of objective and subjective measures can be used to give a good indication of the overall speech quality. A number of objective and subjective measures are therefore presented and several of these measures are then used to evaluate the speech reconstruction algorithms.

5.2 Objective Measures of Speech Quality

The term “objective measures” refer to mathematical expressions that can be used to determine speech quality. Perhaps the most widely known objective measure is the Signal-to-Noise Ratio (SNR). Although SNR works well for the evaluation of PCM speech, its performance is poor for low bit rate speech coders and is mentioned here for the sake of completeness. If $\hat{s}_n$ denotes the original synthetic speech sample and $\hat{s}_{r,n}$ denotes the reconstructed speech sample during frame loss condition, the total SNR can be computed as follows:

$$\text{SNR}_{\text{dB}} = 10 \cdot \log_{10} \left[\frac{\sum_n \hat{s}_n^2}{\sum_n (\hat{s}_{r,n} - \hat{s}_n)^2} \right] \quad (5.1)$$

For the evaluation of speech frame reconstruction algorithms, an SNR is computed for every missing speech frame and then an average SNR value is obtained [27]. Since SNR is sensitive to gain and phase, the speech samples (under comparison) may be normalized and have the time delay removed.

The average SNR is useful for evaluating reconstruction algorithms of PCM speech as a direct comparison can be made between the original and reconstructed waveform samples. However, the SNR measure is less applicable for CELP coded speech because only the CELP speech parameters need to be reconstructed but not the waveform samples. It would be beneficial to make use of objective measures which are immediately related to the CELP parameters. A number of LPC-related measures are discussed in the following sub-sections.

5.2.1 Parametric L_p Distances

For the purpose of developing objective parametric measures, a linear and log L_p norm distance for parametric assessments can be defined respectively as:

$$d_{L_r} = \left[\frac{1}{m} \sum_{i=1}^m |p_{r,i} - p_i|^p \right]^{\frac{1}{p}} \quad (5.2)$$

$$d_{L_r, \text{dB}} = \left[\frac{1}{m} \sum_{i=1}^m |20 \cdot \log |p_{r,i} / p_i||^p \right]^{\frac{1}{p}} \quad (5.3)$$

where p_i and $p_{r,i}$ are the original and reconstructed time domain parameters and m is the order of the parametric model. In particular, Euclidean distances L_2 have received the most attention because of their convenient physical interpretations. Parameters which have been used for this include LPC coefficients, reflection coefficients, LARs, LSPs and cepstral coefficients. The first four parameters have been described previously; the meaning of cepstral coefficients is now explained under the context of log-spectral distance.

5.2.2 Log-Spectral Distance (SD)

The complex cepstrum is obtained by taking the inverse Fourier transform of the log magnitude of the Fourier transform of a windowed speech frame [40]. Based on the linear speech production model for LPC analysis, the short-term spectrum of the speech signal $S(e^{j\omega})$ is the product of a slow-varying vocal tract spectrum $V(e^{j\omega})$ and the fast-varying excitation spectrum $P(e^{j\omega})$:

$$S(e^{j\omega}) = P(e^{j\omega}) V(e^{j\omega}) \quad (5.4)$$

Assuming minimum phase, the logarithm of the speech spectrum can be considered as:

$$\log|S(e^{j\omega})| = \log|P(e^{j\omega})| + \log|V(e^{j\omega})| \quad (5.5)$$

In the additive form, the two components still retain their fast and slow varying behaviour. If they are treated as two “time-varying” signals, the spectral envelope should have “low frequency” contents while the excitation represents the “high frequency” contents. Now by taking the inverse Fourier transform to the log spectrum, an effective transformation is made into one such “frequency” domain known otherwise as quefrency. The cepstral coefficients c_i are then the transform coefficients.

Given a stable LPC synthesis filter $\sigma/A(z)$, the cepstral coefficients can be obtained from the Taylor series of the log magnitude spectrum [40]:

$$\ln[\sigma/|A(e^{j\omega})|^2] = \sum_{k=-\infty}^{\infty} c_k e^{-jk\omega} \quad (5.6)$$

with $c_0 = \ln(\sigma)$ and $c_k = c_{-k}$

Applying the Laurent Expansion to the Taylor series above, a conversion procedure from LPC coefficients to cepstral coefficients can be obtained as below:

$$\begin{aligned} c_1 &= -a_1 \\ c_i &= -a_i - \sum_{k=1}^{i-1} \frac{i-k}{i} c_{i-k} a_k, \quad i = 2, \dots, p \\ c_i &= -\sum_{k=1}^p \frac{i-k}{i} c_{i-k} a_k, \quad i = p+1, \dots, N_c \end{aligned} \quad (5.7)$$

The cepstral coefficients are important because when the cepstral coefficients are used with $p=2$ in (5.2), the distance measure becomes a Euclidean distance between two log magnitude spectra. This can be shown by first defining the log-Spectral Distance (SD) between two LPC filters $1/A_1(z)$ and $1/A_2(z)$ as:

$$d_{SD} = \sqrt{\frac{1}{2\pi} \int_{-\pi}^{\pi} \left| \ln|A_1(e^{j\omega})|^2 - \ln|A_2(e^{j\omega})|^2 \right|^2 d\omega} \quad (5.8)$$

Applying (5.6) to (5.8), the SD can be expressed in terms of cepstral coefficients:

$$d_{SD} = \sqrt{\sum_{k=-\infty}^{\infty} [c_{1,k} - c_{2,k}]^2} = \sqrt{[c_{1,0} - c_{2,0}]^2 + 2 \sum_{k=1}^{\infty} [c_{1,k} - c_{2,k}]^2} \quad (5.9)$$

Since the two LPC filters are normalized to have unity gain such that $c_{1,0} = c_{2,0} = 0$, only the second term in (5.10) is needed to obtain the SD. For this purpose, an objective measure known as Cepstral Distance (CD) [43] is defined:

$$CD_{dB} = \frac{10}{\ln 10} \sqrt{d_{SD}} = \frac{10}{\ln 10} \sqrt{2 \sum_{k=1}^{N_c} [c_{1,k} - c_{2,k}]^2} \quad (5.10)$$

A truncated series of $N_c=30$ terms is usually adequate.

5.2.3 Likelihood Ratio (LR)

The Likelihood Ratio (LR) which was first defined in (4.53) as an alternative error minimization criteria for VLP is often employed as an objective measure in speaker identification applications. In this section, it is shown that the LR can be interpreted as an approximation to the log-spectral distance given in (5.8). By substituting the residual energy expression in (4.51) into (4.54), it is possible to write:

$$d_{LR}(\mathbf{a}', \mathbf{a}'_p) = \frac{\alpha}{\alpha_p} - 1 = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|S(e^{j\omega})|^2 |A(e^{j\omega})|^2}{\alpha_p} d\omega - 1 \quad (5.11)$$

In order to obtain a more recognized form of the above equation, several spectral properties of the LPC filter $\sigma/A(z)$ are used [40]:

1. The log magnitude spectra of $1/A(z)$ has zero mean provided that $1/A(z)$ is stable

and has all of its poles within the unit circle. In other words:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \ln |A(e^{j\omega})|^2 d\omega = 0 \quad (5.12)$$

2. The gain factor $\sigma = \sqrt{\alpha}$ is chosen to match the energy of input speech spectrum $|S(e^{j\omega})|^2$ to the LPC model spectrum $|H(e^{j\omega})|^2$:

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} |S(e^{j\omega})|^2 d\omega = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sigma^2}{|A(e^{j\omega})|^2} d\omega \quad (5.13)$$

3. The average value of the log spectrum of the model can be obtained from (5.14) using (5.13):

$$\int_{-\pi}^{\pi} \ln \left| \frac{\sigma}{A(e^{j\omega})} \right|^2 d\omega = \ln(\sigma^2) \quad (5.14)$$

By Wold's decomposition, the speech and model spectrums are matched exactly as the order of the LPC filter approaches infinity:

$$|S(e^{j\omega})|^2 = \lim_{p \rightarrow \infty} \frac{\alpha_p}{|A_p(e^{j\omega})|^2} \quad (5.15)$$

and by substituting (5.16) into (5.15), the following equation is arrived:

$$\int_{-\pi}^{\pi} \ln |S(e^{j\omega})|^2 d\omega = \ln(\alpha_{\infty}) \quad (5.16)$$

Returning to (5.12) and assuming the speech spectrum $|S(e^{j\omega})|^2$ can be approximated by $\alpha_p/|A_p(e^{j\omega})|^2$ because of (5.15), it is possible to express:

$$d_{LR}(\mathbf{a}', \mathbf{a}'_p) \cong \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{|A(e^{j\omega})|^2}{|A_p(e^{j\omega})|^2} d\omega - 1 \quad (5.17)$$

Using the Taylor series approximation where:

$$u = e^{(\ln(u))} \approx 1 + \ln(u) + \frac{1}{2} (\ln(u))^2 \quad u = 1 \quad (5.18)$$

and making use of (5.13), (5.17) then becomes:

$$\begin{aligned}
 d_{LR}(a', a'_p) &\equiv 1 + \frac{1}{2\pi} \int_{-\pi}^{\pi} \ln \left(\frac{|A(e^{j\omega})|^2}{|A_p(e^{j\omega})|^2} \right) d\omega \\
 &\quad + \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\ln \left(\frac{|A(e^{j\omega})|^2}{|A_p(e^{j\omega})|^2} \right) \right)^2 d\omega - 1 \\
 d_{LR}(a', a'_p) &\equiv \frac{1}{4\pi} \int_{-\pi}^{\pi} (\ln |A(e^{j\omega})|^2 - \ln |A_p(e^{j\omega})|^2)^2 d\omega
 \end{aligned} \tag{5.19}$$

Thus for very small distortion, the LR is roughly proportional to the SD defined in (5.8).

5.2.4 Other Objective Distortion Measures

There are a number of other spectral distortion measures which are frequently used to judge the speech quality. Based on the norm L_p parametric distance definition, a frequency weighted log measure [43] which gives more weight to distortion in the regions of greater spectral energy can be defined:

$$D_w = \left\{ \frac{\sum_{k=0}^{L-1} P_x(e^{j2k(\pi/L)}) \cdot \left| 10 \log \left[\frac{P_x(e^{j2k(\pi/L)})}{P_y(e^{j2k(\pi/L)})} \right] \right|^p}{\sum_{k=0}^{L-1} P_x(e^{j2k(\pi/L)})} \right\}^{\frac{1}{p}} \tag{5.20}$$

where P_x and P_y are the respective reference power spectra and particular power spectra. L is the number of frequency points and p is the distance norm.

A related distortion method known as the Bark Spectral Distortion (BSD) [60] takes note of the nonlinear sensitivity of the human auditory system to frequency and amplitude. The speech is processed through a filter bank and then individual frequency bands are weighted differently to adjust for frequency dependent hearing sensitivity. The BSD between the n^{th} original and reconstructed speech frame can be defined as:

$$BSD^{(n)} = \sum_{i=1}^N [S^{(n)}(i) - S_r^{(n)}(i)]^2 \quad (5.21)$$

where N is the number of critical frequency bands and $S(i)$ and $S_r(i)$ are the weighted spectrum of the original and reconstructed speech.

Another class of distortion measures has been developed based on the use of squared coherency function [43] defined as:

$$\gamma^2(e^{j\omega}) = \frac{|P_{xy}(e^{j\omega})|^2}{P_x(e^{j\omega})P_y(e^{j\omega})} \equiv \frac{\left| \sum_n X_n^*(e^{j\omega}) Y_n(e^{j\omega}) \right|^2}{\sum_n |X_n(e^{j\omega})|^2 \sum_n |Y_n(e^{j\omega})|^2} \quad (5.22)$$

where $X_n(e^{j\omega})$ and $Y_n(e^{j\omega})$ are the frequency spectra of the n^{th} speech frames of the two signals in comparison. The coherence function $\gamma^2(e^{j\omega})$ acts as a correlation coefficient for a specific angular frequency ω . Utilizing the squared coherency function, a Signal-to-Distortion Ratio (SDR) measure can be computed as follows:

$$SDR(e^{j\omega})_{\text{dB}} = 10 \cdot \log \left(\frac{\gamma^2(e^{j\omega})}{1 - \gamma^2(e^{j\omega})} \right) \quad (5.23)$$

5.3 Subjective Measures of Speech Quality

Although the aforementioned objective measures are useful in the assessment of the speech reconstruction algorithms, the speech quality should be judged ultimately by subjective listening tests.

In general, subjective tests are concerned with two aspects of speech, namely intelligibility and quality. Although good quality always implies good intelligibility, the opposite may not be true especially in the case of low bit-rate speech coder design. Since the main objective of this study is to evaluate the reconstructed speech quality during frame loss conditions, only quality tests are discussed.

5.3.1 A-B Preference Test

Known otherwise as the paired comparison test, this is one of several formal listening tests recommended by CCITT [6]. In the A-B preference test, a sentence pair is processed

by two different algorithms and the sentence pair is presented to the listener in a random order. The listener then decides if sentence A or sentence B is preferred. The "no difference" response is not allowed in this test. In order to ensure the reliability of the listener, the sentence pairs may be repeated for consistency and reference condition such as original uncorrupted speech can be used. Result of the paired comparison test is usually presented as a percentage of preference for a particular algorithm.

5.3.2 Mean Opinion Score (MOS) Test

Unlike the A-B test which provides only a binary decision on the listening test results, the Mean Opinion Score (MOS) test enables subjective ranking of a particular algorithm among other ones. In the MOS test, a sentence pair belonging to the same speaker is picked from a speech database and then processed through a randomly picked algorithm (with a particular test condition). The listeners are asked to rank the speech on an a scale between 1 and 5 [57][6]. Since MOS is not an absolute measure, standard reference conditions should be inserted to anchor the listeners and facilitate comparisons across different tests.

5.3.3 Other Subjective Measures

A thorough quality assessment can be made through the use of the Diagnostic Acceptability Measure (DAM) [44]. In the DAM, highly-trained listeners are presented with recordings taken from a list of phonetically balanced sentences and they are asked to rate the speech quality both in terms of overall acceptability as well as parametric characteristics. For certain predefined characteristics, the listeners would rate the qualities with respect to the background, the speech (itself) and the total effect.

Results of these ratings are processed with different weightings depending upon the perceptual importance of individual parameters. The final ratings are used in a multiple, non-linear regression equation to give a parametric estimate of the acceptability of the algorithm.

5.4 Evaluation of Speech Reconstruction Schemes

Utilizing the parameter classification approach described in Section 4.2.2, the

reconstruction schemes were designed and simulated (in C language) for various FERs. For a particular ICELP parameter, reconstructed frames which had the worst distortions were found by the following means:

1. Informal listening test on the reconstructed speech and visual inspection of original and reconstructed speech waveforms;
2. Objective measures based on L_2 distance for LPC coefficients (a_n), reflection coefficients (k_n), LSPs (ω_n), frame gain (f_n) and stochastic excitation gain values (g_n), CD and LR for the LPC coefficients (a_n) and
3. Comparison of original and reconstructed time series of the particular ICELP parameter under examination.

In this way, the offending parameter could be isolated quickly and analyzed to enable suitable refinements to be made to the reconstruction method. Table 5.1 summarizes some typical average results obtained for the ICELP short-term LPC parameters.

Reconstruction Methods	Average L_2 k_n/ω_n	Average L_2 a_n	Average CD (dB)	Average LR
Repeat-K	0.18	0.24	3.83	1.08
MSE-K-VLP-1S	0.17	0.24	4.27	1.19
MSE-K-VLP-2S	0.14	0.24	3.26	0.62
LR-K-VLP-1S	0.17	0.23	3.70	0.77
LR-K-VLP-2S	0.14	0.19	2.92	0.41
Repeat-LSP	0.026	0.24	3.91	1.48
MSE-LSP-VLP-1S	0.024	0.23	3.57	0.76
MSE-LSP-VLP-2S	0.018	0.23	2.77	0.37
LR-LSP-VLP-1S	0.024	0.23	3.68	0.75
LR-LSP-VLP-2S	0.019	0.18	2.90	0.39

Table 5.1 - Distortion Measurements for Short-Term LPC Parameters

Based on the results obtained from the objective measurements and informal listening tests, the following observations are made:

- For MSE-VLP reconstruction schemes, the average SDs of LSPs are always lower than those obtained for the reflection coefficients. Similar findings have also been reported in [64] where LSPs were used for interframe LPC in low bit rate speech coding. The LSPs appear to be better parameters for parametric extrapolation and interpolation.
- Little is gained in using MSE-K-VLP-1S over the simple last frame repetition method. Although the MSE of the reconstructed reflection coefficients is less than the MSE of last frame repetition method, the SD may be worse as revealed in the tabulated results. Informal listening tests indicated little improvement was made when MSE-K-VLP-1S was used for reconstruction.
- Interpolation (VLP-2S) is better than extrapolation (VLP-1S) because of the availability of future information. The CDs of the interpolation methods are approximately 0.6 dB lower than the extrapolation while almost a factor of 2 is noted in the case of the LRs. A noted improvement could be heard for the reflection coefficients while improvement is less marked in the case of LSPs.
- Use of the LR as a minimization criteria yields the best objective and subjective results for reflection coefficients. The spectral distortions are smaller than the reconstruction based on minimizing the MSE. On the other hand, listening tests indicate the LSPs are less sensitive to either minimization criterion. Although the LR-LSP-VLP schemes are perceptually better, the objective spectral distortion measurements were slightly worse than MSE-LSP-VLP as shown in Table 5.1. A plausible explanation is that some of the reconstructed LPC coefficients might have resulted in unstable filter and hence the subsequent stabilization led to increased spectral mismatches between the original and synthesized speech.

In order to benchmark the subjective performance of both the extrapolation and interpolation methods, the VSELP coder which has similar bit rate and performance to ICELP can be employed. The IS-54 bad frame masking procedures described in Section 3.3.3.2 were employed. Furthermore, an “enhancement” was made whereby the VSELP subframe parameters (which include GS-P0-P1 gain codes, pitch lags and stochastic codebook indices) were repeated from the most recent correctly received frames during frame loss conditions. Suitable “anchoring” conditions: Modulated Noise Reference Units

(MNRUs) [7] and error-free ADPCM which have reliable results from other listening experiments [35][57] were also employed in the subjective tests. In order to keep the number of test conditions to a manageable size, only the LR-LSP-VLP-1S and LR-LSP-VLP-2S reconstruction methods were chosen for the ICELP coder.

For this particular study, 19 test conditions which include 3 MNRU values at 10dB (MNRU10), 15dB (MNRU15) and 20dB (MNRU20) and 16 other conditions at various FERs (as shown in Table 5.2 below) were employed.

FER	IS-54 VSELP + Bad Frame Masking	ICELP + LR-LSP-VLP-1S	ICELP + LR-LSP-VLP-2S	G.721 ADPCM
0%	VSELP0	ICELP0	ICELP0	ADPCM0
1%	VSELP1	ICELP1	ICELP2S1	-
2%	VSELP2	ICELP2	ICELP2S2	-
5%	VSELP5	ICELP5	ICELP2S5	-
10%	VSELP10	ICELP10	ICELP2S10	-

Table 5.2 - MOS Test Conditions at 0%, 1%, 2%, 5% and 10% FERs

The source material consisted of 224 sentence pairs (encompassing 112 male-spoken and 112 female-spoken recordings) extracted from the TIMIT speech database. Each sentence pair contained approximately 10 seconds of speech recording and was IRS-filtered to simulate typical channel conditions in the PSTN. After the IRS-filtering, the sentence pairs were stored in three separate formats: 16-bit linear, compressed-VSELP and compressed-ICELP to enable real-time speech decoding during listening tests. During a particular MOS test session, the listeners were asked to rate 91 separate recordings each with a random test condition assigned from one of the above and applied to a random sentence pair. The MOS test procedures (Appendix C) were devised to adhere closely to the CCITT Recommendation [6]. The scores were entered manually by the listeners on specially prepared score sheets. In total, 3480 test cases equivalent to roughly 10 hours of speech were presented to 38 test subjects (27 males and 11 females).

On completion of the listening tests, the results were entered into a computer for statistical analysis. The MOS and error margins (based on 95% confidence intervals) were

obtained for each of the 19 test conditions. The MOS and associated error margins of male, female and combined speech are shown in Table 5.3 and further displayed as a bar chart in Fig. 4.13.

Test Conditions	Mean Opinion Score (MOS)			Error (95% confidence limit)		
	Combined	Male	Female	Combined	Male	Female
MNRU10	1.92	2.02	1.80	0.12	0.15	0.17
MNRU15	3.07	3.23	2.92	0.14	0.20	0.19
MNRU20	3.88	4.11	3.64	0.15	0.18	0.23
ADPCM0	4.02	4.01	4.04	0.14	0.19	0.21
VSELP0	3.81	3.80	3.83	0.15	0.19	0.23
VSELP1	3.62	3.68	3.55	0.15	0.19	0.24
VSELP2	3.40	3.40	3.39	0.14	0.22	0.18
VSELP5	2.97	3.01	2.93	0.14	0.20	0.18
VSELP10	2.54	2.54	2.54	0.12	0.18	0.17
ICELP0	3.73	3.60	3.83	0.13	0.20	0.16
ICELP1	3.72	3.75	3.68	0.13	0.16	0.21
ICELP2	3.51	3.49	3.53	0.13	0.18	0.18
ICELP5	3.13	3.28	3.02	0.15	0.24	0.19
ICELP10	2.59	2.53	2.66	0.13	0.18	0.20
ICELP2S1	3.70	3.58	3.83	0.12	0.18	0.17
ICELP2S2	3.63	3.62	3.63	0.14	0.17	0.21
ICELP2S5	3.24	3.30	3.20	0.14	0.23	0.18
ICELP2S10	2.68	2.63	2.74	0.14	0.18	0.21

Table 5.3 - MOS of Reconstruction Methods and Reference Conditions

The MOS test results and listeners' comments are summarized in the following list:

- The combined MOS of the three MNRU conditions agree quite well with the results reported in [57] whereas the combined MOS of error-free ADPCM is also close to the 4.10 value in [29]. Since the subjective results of both the MNRUs and ADPCM compare favourably with the ones previously reported in the literature, the subjective results of the various reconstruction methods could be judged reliably with respect to these reference conditions.
- Under the no loss condition, the VSELP coding scheme appears to perform slightly better than the ICELP coder in the male speech. The "modified" pitch search

strategy employed in the ICELP coder design might have impacted the speech quality. However, considering the magnitude of the error margins, it is relatively difficult to say this with certainty.

- Despite some overlap of the error margins, the VLP-based extrapolation and interpolation methods perform consistently better than the VSELP method at 1%, 2%, 5% and 10% FERs as indicated in Fig. 4.13. It is important to note that there is *almost* no degradation in MOS for the ICELP coder when the FER increases from 0% to 1% whereas the VSELP MOS is reduced by 0.21. At 2% FER, the VSELP MOS has almost dropped by 0.41 while VLP-1S and VLP-2S reduces the amount of MOS degradation to 0.22 and 0.10 respectively. At 5% FER, the VSELP MOS has dropped below 3.0 while the MOS of ICELP are still above 3.0. The gaps between the MOS of the three methods become much closer when the FER is at 10%.
- An equally-weighted composite MOS (MOS_c) [57] is obtained for the three reconstruction schemes using the following equation:

$$MOS_c = \frac{\sum_{i=0,1,2,5,10} MOS_i \cdot n_i}{\sum_{i=0,1,2,5,10} n_i} \quad (5.24)$$

where MOS_i and n_i denote respectively the MOS and sample size at $i\%$ FER. Confidence intervals (95%) are also calculated for the respective MOS values. The results (in Table 5.4) are depicted as a bar chart with error margins in Fig. 4.14. It can be observed the composite MOS values of the extrapolation and interpolation methods are respectively 0.10 and 0.12 higher than the VSELP method.

Method	MOS_c	Error (95% C.L.)
ICELP + LR-LSP-VLP-1S	3.35	0.064
ICELP + LR-LSP-VLP-2S	3.37	0.062
VSELP + Frame Masking	3.25	0.065

Table 5.4 - Composite MOS of Reconstruction Methods

- During the MOS tests, the listeners found the reconstructed speech with pops and

Figure 4.13 - MOS Results of Various FER and Test Conditions

Figure 4.14 - Composite MOS Results

clicks to be most disturbing perceptually. These were most likely caused by energy mismatch between the original and reconstructed gain parameters. At a less disturbing level, several listeners also noticed a faint whistling sound and word stuttering in the reconstructed ICELP coded speech on some occasions. It would appear the former effect was caused by errors in the reconstructed LSPs as noted in [10] whereas the latter could be due to incorrect estimation of the ACB parameters, notably the pitch lag and ACB coefficients in the lost frames.

5.5 Summary

In this chapter, a number of objective and subjective quality measures suitable for use in the evaluation of speech frame reconstruction algorithms were described. To facilitate evaluation of the speech frame reconstruction methods, an approach combining objective and subjective measurements was employed. In the initial development stage, the L_2 norm distance, CD and LR measures were computed for the ICELP parameters and refinements were made to the reconstruction algorithms in an iterative manner until the reconstructed speech quality was satisfactory. Formal subjective listening tests were then performed for several reconstruction algorithms for the ICELP coder. The IS-54 VSELP speech coder employing the frame substitution method was also included as a reference in the test.

Chapter 6

Conclusions

6.1 Conclusions

In this research, speech coding over the frame relay and digital cellular networks was reviewed and the frame loss problem common to both applications was addressed for the CELP speech coders. Since the CELP algorithm is parametric in nature, it is feasible to estimate the lost frame parameters from those already received. Parametric extrapolation and interpolation methods utilizing VLP were therefore presented for speech frame reconstruction.

The major conclusions of this research study are summarized below:

- Based on both objective and subjective measurement results, the 2-sided VLP method (interpolation) was always better than the 1-sided VLP method (extrapolation) because of the availability of future information. However, the improved performance came at the expense of added buffering delay (equivalent to a single speech frame).
- When compared to the MSE-VLP methods, the LR-VLP methods were found to work best for estimating the short-term LPC parameters. In particular, LSPs were shown to have better reconstruction properties than the reflection coefficients. The difference in performance between MSE-LSP-VLP and LR-LSP-VLP was minimal and MSE-LSP-VLP might be used to avoid extra parametric transformation.
- In most cases, the frame gain and stochastic excitation gain parameters could be

reconstructed (with a fair degree of accuracy) using VLP based on minimizing the MSE criterion.

- Except for the ACB coefficient b_0 , the remaining ACB parameters, namely the pitch lag and ACB coefficients $\{b_{-1}, b_1\}$, were not suitable for VLP because of erratic behaviour in their time series. A modified pitch search strategy was employed instead to “smooth” the pitch lag contour during speech encoding. Median filtering was then applied at the decoder to estimate the lost pitch lags. A conditional repeat strategy was employed for reconstructing the ACB coefficients.
- The stochastic codebook indices and sign multipliers may be chosen randomly without significant degradation in perceptual quality.
- An adaptive energy control mechanism was devised to ensure there was no substantial energy increase in reconstructed speech frames when compared to the previously received frame.
- Subjective MOS test results indicated the extrapolation and interpolation methods were consistently better than the IS-54 VSELP bad frame masking strategy for random FERs up to 10%. At 1% FER, the ICELP coder suffered almost no degradation in perceptual quality when compared to the error-free condition.
- Whereas the extrapolation method is more suitable for use in the digital cellular radio applications (for minimal processing delay), the interpolation method can be useful in the frame relay applications as it is possible for “future” speech frames to arrive ahead of the lost frame due to variable network delay and buffering in the adaptive buildout delay mechanism.
- The extrapolation and interpolation methods are simple to implement as it amounts only to a slight increase in computational complexity mainly due to the adaptive energy control mechanism. Since the reconstruction method only makes use of the speech parameters, there is no overhead in the speech frame transmission except for an error detection mechanism such as CRC which already exists in both the frame relay and digital cellular applications.
- As part of the adaptive buildout delay mechanism employed in the frame relay

networks, a simple rate adjustment algorithm combining fixed threshold VAD and TSM was shown to be feasible for use with the ICELP coder. Specifically, the TSM scheme is based on replication or truncation of the codebook excitation sequence. It was observed that degradation in voice quality could be minimized if large adjustments were made only in the silent frames and if the number of time-scale modified frames was kept to no more than 2% of the total number of frames.

6.2 Suggestions for Future Research

This final section is divided into two parts. In the first part, a number of refinements to the existing frame relay speech coder design are proposed. In the second part, recommendations pertaining to speech frame reconstruction are presented.

6.2.1 Frame Relay Speech Coder

The delay equalization method employed by the adaptive buildout delay mechanism was briefly examined in this study. Formal MOS results could be obtained to determine the exact percentage of time-scale modified frames without incurring major degradation in output speech quality. At the same time, improved VAD methods other than fixed threshold energy detection could be investigated for transmission of noisy speech frames.

In the current fixed rate 8 kb/s ICELP coder, any speech frame (regardless of its content) can be dropped in the frame relay network because of traffic congestions. To minimize the loss of speech frames that are hard to reconstruct, an "analysis-by-synthesis" frame reconstruction method could be applied at the ICELP encoder. Each encoded speech frame is compared with its reconstructed counterpart. If there is only a small difference between the two frames, the DE bit of the frame relay frame could be set accordingly to denote the particular frame could be dropped without seriously affecting the speech frame reconstruction at the receiving end.

A 16 kb/s coder which consists of an embedded 8 kb/s CELP coder and an auxiliary 8 kb/s multi-pulse excitation was outlined in this thesis. By simulating suitable frame relay network conditions, MOS results of this 16 kb/s MPE-CELP coder could be examined and compared to the fixed rate 8 kb/s ICELP coder supplemented with the setting of DE bit as well as the speech frame reconstruction methods described earlier.

6.2.2 Speech Frame Reconstruction

Although the VLP method works quite well in reconstructing several of the ICELP parameters, the following areas could be explored to improve the existing extrapolation and interpolation schemes:

- LAR is used as an alternative LPC parameter in the RPE-LTP coder design and has shown good properties for linear interpolation. Use of LAR for parametric reconstruction based on either MSE-VLP or LR-VLP could be investigated and compared with the results obtained for the reflection coefficients and LSPs.
- In order to further reduce the amount of spectral distortion between the original and reconstructed short-term LPC parameters, other distortion minimization criteria suitable for use with the VLP method could be examined.
- It was found that VLP was less effective in reconstructing the ACB parameters. The fractional rate long-term LP filter which was reported to perform better than the conventional multi-tap ACB could be introduced into the ICELP design. The use of a single gain term and a fractional pitch lag may lead to better reconstruction of the lost values. Additionally, non-linear prediction techniques may be worth examining in the reconstruction of ACB parameters.
- It was shown in Sections 4.2.2.1 and 4.2.2.4 that large errors may arise occasionally in the frame gain and stochastic excitation gain estimates (especially for small gain values) and hence lead to energy explosion in the reconstructed speech. The dynamic range of the original values could be reduced to enable better extrapolation and interpolation of the gain parameters by performing the VLP in the logarithmic domain [8].
- Since the adaptive energy control mechanism can only compute the subframe energy directly from the synthesized speech waveform, extra computational cost and processing delay are needed to resynthesize the CELP coded speech if the current subframe energy exceeds the value in the previous subframe. A better approach is to find a way to perform the energy check in the parametric domain prior to speech synthesis and suppress sudden energy increase in the reconstructed speech.

- Real-time implementation of the ICELP coder coupled with the speech frame reconstruction methods could be examined. First order VLP-IS method may have to be used instead to allow for possible trade-offs in computational complexity.
- The VLP coefficients are fixed in the current extrapolator and interpolator design. Backward adaptation could be used to provide continuous update of the VLP coefficients. Another approach is to design two sets of VLP functions based on silence and non-silence (or silence, voiced and unvoiced) frame classification similar to the ones described in [64][12].
- A random frame loss condition is currently employed to evaluate the speech frame reconstruction methods. It is suggested queueing and voice activity simulation of either the frame relay network [13] or general packet network [65] could be used for more realistic modelling of the frame loss statistics. For digital cellular systems, a channel model which consists of Rayleigh fading and additive Gaussian white noise could be used to generate the frame loss distribution at various carrier/interference ratios and vehicular speeds [54][18].
- A much larger population size of 60-100 listeners should be employed in future MOS tests to ensure proper separation of the error margins of the test conditions.

References

- [1] M.K. Asi and B.E.A. Saleh, "A Linear Periodically Time-Varying Filter for Time-Frequency Scaling of Speech," *Proc. of IEEE ICASSP*, pp. 405-408, 1990.
- [2] J.P. Campbell, Jr., T.E. Tremain and V. C. Welch, "The DOD 4.8 KPS Standard (Proposed Federal Standard 1016)," *Advances in Speech Coding*, ed. by B.S. Atal *et. al.*, pp. 121-134, Kluwer Academic Publishers, Norwell, Massachusetts, 1991.
- [3] J.P. Campbell, Jr., V.C. Welch and T.E. Tremain, "An Expandable Error-Protected 4800 bps CELP Coder," *Proc. IEEE ICASSP*, pp. 735-738, 1989.
- [4] J. V. Candy, *Signal Processing - the Modern Approach*, McGraw-Hill Book Company, New York, New York, 1988.
- [5] J.P. Cavanagh, "Applying the Frame Relay Interface to Private Networks," *IEEE Communications Magazine*, vol.30, No. 3, pp.48-64, March 1992.
- [6] CCITT, Supplement No. 14, "Subjective Performance Assessment of Digital Processes using the Modulated Noise Reference Unit (MNRU)," *Blue Book*, Vol. V, 1988.
- [7] CCITT, Recommendation P.81, "Modulated Noise Reference Noise Unit (MNRU)," *Blue Book*, Vol. V, 1988.
- [8] J.H. Chen. "A Robust Low-Delay CELP Speech Coder at 16 kb/s," *Advances in Speech Coding*, ed. by B.S. Atal *et. al.*, pp. 25-35, Kluwer Academic Publishers, Norwell, Massachusetts, 1991.
- [9] J.H. Chen and A. Gersho, "Covariance and Autocorrelation Methods for Vector Linear Prediction," *Proc. of IEEE ICASSP*, pp. 1545-1548, 1987.
- [10] R.V. Cox, W. Bastiaan Kleijn and P. Kroon, "Robust CELP Coders for Noisy Background and Noisy Channels," *Proc. IEEE ICASSP*, pp.739-740, 1989.
- [11] R.E. Crochiere and L.R. Rabiner, *Multirate Digital Signal Processing*, Prentice-

- Hall Inc., Englewood Cliffs, New Jersey, 1983.
- [12] V. Cuperman and A. Gersho, "Vector Predictive Coding of Speech at 16 kbits/s," *IEEE Trans. on Communications*, Vol. 33, No. 7, pp. 685-696, July 1985
 - [13] L. Dong, *Compressed Voice in Integrated Services Frame Relay Network*, M.Eng. Thesis, Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Nov. 1992.
 - [14] EIA Interim Standard IS-54, *Cellular System: Dual-Mode Subscriber Equipment - Network Equipment Compatibility Specification*, U.S.A., December 1989.
 - [15] P.M. Embree and B. Kimble, *C Language Algorithms for Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, New Jersey, 1991.
 - [16] ETSI/GSM Recommendation 06.10, *European Digital Cellular Telecommunication System (Phase 1): Full-rate Speech Transcoding*, France, March 1992.
 - [17] ETSI/GSM Recommendation 06.11, *Substitution and Muting of Lost Frames for Full Rate Speech Traffic Channels*, France, Feb. 1992.
 - [18] C.G. Gerlach, "A Probabilistic Framework for Optimum Speech Extrapolation in Digital Mobile Radio," *Proc. of IEEE ICASSP*, Vol. II, pp. 419-422, 1993.
 - [19] A. Gersho and R.M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, Massachusetts, 1992.
 - [20] I.A. Gerson and M.A. Jasiuk, "Vector Sum Excited Linear Prediction (VSELP) Speech Coding at 8 kbps," *Proc. of IEEE ICASSP*, pp. 461-464, 1990.
 - [21] D.J. Goodman, "Second Generation Wireless Information Networks," *IEEE Trans. on Vehicular Technology*, Vol. 40, No. 2, pp. 366-374, May 1991.
 - [22] D.J. Goodman, "Embedded DPCM for Variable Bit Rate Transmission," *IEEE Trans. on Communications*, Vol. 28, No. 7, pp. 1040-1046, July 1980.
 - [23] D.J. Goodman, "Trends in Cellular and Cordless Communications," *IEEE Communications Magazine*, vol. 29, No. 6, pp.31-40, June 1991.
 - [24] D.J. Goodman, G.B. Lockhart, O.J. Wasem and W.C. Wong, "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 34, pp. 1440-1448, Dec. 1986.
 - [25] J.G. Gruber and L. Strawczynski, "Subjective Effects of Variable Delay and Speech Clipping in Dynamically Managed Voice Systems," *IEEE Trans. on*

Communications, Vol. 33, No. 8, pp. 801-808, August 1985.

- [26] M.R.L. Hodges, "The GSM Radio Interface," *British Telecom Technical Journal*, Vol. 8, No. 1, pp. 31-43, January 1990.
- [27] O.G. Jaffe, *Reconstruction of Missing Packets of PCM and ADPCM Encoded Speech*, Master's Thesis, Massachusetts Institute of Technology, Cambridge, Massachusetts, June 1986.
- [28] N.S. Jayant and S.W. Christensen, "Effects of Packet Losses in Waveform Coded Speech and Improvements due to an Odd-Even Sample-Interpolation Procedure," *IEEE Trans. on Communications*, Vol. 29, No. 2, February 1981, pp.101-109.
- [29] N.S. Jayant, "High Quality Coding of Telephone Speech and Wideband Audio," *IEEE Communications Magazine*, Vol. 28, No. 1, pp.10-20, January 1990.
- [30] P. Kabal and R. Ramachandran, "The Computation of Line Spectral Frequencies Using Chebyshev Polynomials," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 34, No. 6, Dec. 1986.
- [31] N. Kitawaki *et. al.*, "Speech Coding Technology for ATM Networks," *IEEE Communications Magazine*, Vol. 28, No. 1, pp. 21-27, January 1990.
- [32] P. Kroon and B.S. Atal, "On Improving the Performance of Pitch Predictors in Speech Coding Systems," *Advances in Speech Coding*, ed. by B.S. Atal *et. al.*, pp. 311-320, Kluwer Academic Publishers, Norwell, Massachusetts, 1991.
- [33] P. Kroon and E.D. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s," *IEEE Journal on Selected Areas in Communications*, vol. 6, No. 2, pp. 353-363, February 1988.
- [34] W.P. LeBlanc, *An Advanced Speech Coder based on A Rate-Distortion Theory Framework*, M.Eng. Thesis, Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, March 1988.
- [35] W.P. LeBlanc, *Speech Coding at Low to Medium Bit Rates*, Ph.D. Thesis, Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, October 1992.
- [36] W.P. LeBlanc, V. Cuperman, B. Bhattacharya and S.A. Mahmoud, "Efficient Search and Design Procedures for Robust Multi-Stage VQ of LPC Parameters for 4 kb/s Speech Coding," *IEEE Trans. on Speech and Audio Processing*, Oct. 1992.
- [37] T.W. Leung, W.P. LeBlanc, S.A. Mahmoud, "Speech Coding over Frame Relay Networks," *Proc. of IEEE Workshop on Speech Coding*, Sainte-Adèle, Quebec,

Oct. 1993.

- [38] G.B. Lockhart and M.M. Lara-Barron, "Implementation of Packet-Based Encoding Schemes for Speech Transmission," *Proc. of 6th IEE Int. Conference on Digital Processing of Signals in Communications*, pp. 326-330, Loughborough, U.K., 1991.
- [39] J. Makhoul, S. Roucos and H. Gish, "Vector Quantization in Speech Coding," *Proc. of IEEE*, vol. 73, No. 11, pp. 1551-1588, Nov. 1985.
- [40] J.D. Markel and A.H. Gray, Jr., *Linear Prediction of Speech*, Springer-Verlag, New York, New York, 1976.
- [41] W.A. Montgomery, "Techniques for Packet Voice Synchronization," *IEEE Jou. on Selected Areas in Communications*, vol. 1, No. 6, pp. 1022-1028, December 1983.
- [42] A. M. Noll, "Voice vs Data: An Estimate of Future Broadband Traffic," *IEEE Communications Magazine*, vol. 29, No. 6, pp.22-29, June 1991.
- [43] "Objective Assessment of Voice Quality," Contribution to Working Group T1Y1.2, Doc. No. T1Y1.2/91-083, July 1991.
- [44] P.E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice-Hall, Englewood Cliffs, New Jersey, 1987
- [45] M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," *IEEE Trans. on ASSP*, vol. 29, No. 3, pp. 374-390, June 1981.
- [46] J.G. Proakis, *Digital Communications 2nd Ed.*, McGraw-Hill Book Company, New York, New York, 1989.
- [47] T.F. Quatieri and R.J. McAulay, "Speech Transformations Based on a Sinusoidal Representation," *IEEE Trans. on ASSP*, vol. 34, No. 6, pp. 1449-1464, Dec. 1986.
- [48] L.R. Rabiner, M.R. Sambur and C.E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 23, No. 6, pp. 552-557, Dec. 1975.
- [49] K. Raith and J. Uddenfeldt, "Capacity of Digital Cellular TDMA Systems," *IEEE Trans. on Vehicular Technology*, Vol. 40, No. 2, pp. 323-332, May 1991.
- [50] R.P. Ramachandran and P. Kabal, "Stability and Performance Analysis of Pitch Filters in Speech Coders," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-35, No. 7, pp. 937-946, July 1987.
- [51] M.R. Schroeder and B.S. Atal, "Code-Excited Linear Prediction (CELP): High

- Quality Speech at Very Low Bit Rates," *Proc. of IEEE ICASSP*, pp. 937-940, 1985.
- [52] C. B. Southcott *et. al.*, "Voice Control of the Pan-European Digital Mobile Radio Systems," *Proc. IEEE ICASSP*, pp. 1070-1074, 1989.
 - [53] G.W. Stewart, *Introduction to Matrix Computation*, Academic Press, Inc., Orlando, Florida, 1973.
 - [54] H.-Y. Su and P. Mermelstein, "Improving The Speech Quality of Cellular Mobile Systems under Heavy Fading," *Proc. of IEEE ICASSP*, Vol. II, pp. 121-124, 1992.
 - [55] J. Suzuki and M. Taka, "Missing Packet Recovery Techniques for Low Bit Rate Coded Speech," *IEEE J.S.A.C.*, Vol. 7, No. 5, pp. 707-717, June 1989.
 - [56] J. Tang, "Evaluation of Double Sided Periodic Substitution (DSPS) Method for Recovering Missing Speech in Packet Voice Communication," *Proc. of 10th Annual IEEE Int. Phoenix Conference on Computers and Communications*, pp. 454-458, 1991.
 - [57] L.A. Thorpe and P.V. Coverdale, "Evaluation of Digital Cellular Speech Coders: Listening Test Results," *Bell Northern Research Report* TR-4183.
 - [58] P. Vary *et. al.*, "Speech Codec for the European Mobile Radio System," *Proc. of IEEE Globecom.*, Dallas, November 1989.
 - [59] R.A. Valenzuela and C.N. Animalu, "A New Voice-Packet Reconstruction Scheme," *Proc. of IEEE ICASSP*, pp. 1334-1336, 1989.
 - [60] S. Wang, A. Sekey and A. Gersho, "Auditory Distortion Measure for Speech Coding," *Proc. of IEEE ICASSP*, pp. 493-496, 1991.
 - [61] O.J. Wasem, D.J. Goodman, C.A. Dvorak and H.G. Page, "The Effect of Waveform Substitutions on the Quality of PCM Packet Communications," *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol. 36, March 1988.
 - [62] C.J. Weinstein and J.W. Forgie, "Experience with Speech Communications in Packet Networks," *IEEE Jou. on Selected Areas in Communications*, Vol. 1, No. 6, pp. 963-980, December 1983.
 - [63] R. A. Wiggins and E.A. Robinson, "Recursive Solution to the Multichannel Filtering Problem," *Journal of Geophysical Research*, vol. 70, no. 8, pp.1885-1891, April 1965.
 - [64] M. Yong, G. Davidson and A. Gersho, "Encoding of LPC Spectral Parameters using Switched-Adaptive Interframe Vector Prediction," *Proc. of IEEE ICASSP*, pp. 402-405, 1988.

- [65] M. Yong, "Study of Voice Packet Reconstruction Methods Applied to CELP Speech Coding," *Proc. of IEEE ICASSP*, pp. II-125-128, 1992.
- [66] S. Zhang and A.R. Kaye, Private Communications, October 1992.

Appendix A

Sampling Rate Conversion

The process of sampling rate reduction is called: *Decimation*. An $M:1$ decimation refers to an M -fold reduction in sampling frequency f_s such that for every M input samples (at a rate of f_s), a decimated output (at a rate f_s/M) is obtained. The process of sampling rate increase is called *Interpolation*. An $1:L$ interpolation refers to an L -fold increase in sampling frequency such that L interpolated samples are generated (at a rate of Lf_s) for every input sample (at a rate of f_s). Likewise, an $M:L$ *Fractional-Rate Conversion* denotes a non-integral change in sampling rate. Thus for every block of M input samples (at a rate of f_s), L output samples (at a rate of Lf_s/M) are obtained. Nyquist criterion must be observed by all sampling rate conversion processes.

Referring to Fig. 2.10(a), the $M:1$ decimated output $y(m)$ can be written as a convolution sum between the input $x(n)$ and the filter impulse response $h(n)$:

$$y(m) = \sum_{k=-\infty}^{\infty} h(Mm - n) x(n) \quad (\text{A.1})$$

The digital filter serves the purpose of an anti-aliasing filter and has a frequency response $H(e^{j\omega})$ which approximates the following ideal lowpass characteristics:

$$\tilde{H}(e^{j\omega}) = \begin{cases} 1, & |\omega| \leq \frac{2\pi f_s}{2M} = \frac{f_s}{M} \\ 0, & \text{otherwise} \end{cases} \quad (\text{A.2})$$

The filtered signal $w(n)$ at a sampling rate of f_s is therefore fed into a rate compressor

which outputs every M^{th} sample of input $w(n)$ such that $y(m) = w(Mm)$. The index variable is altered from n to m in (A.1) to denote the output $y(m)$ is now generated at a rate of f_s/M .

For the interpolator shown in Fig. 2.10(b), $L-1$ zero-valued samples are inserted between every input sample at the $1:L$ sample rate expander. Since this has the effect of replicating the baseband spectrum at $\pm n f_s$, the aliased spectral images must be removed from the output $w(m)$. Similar to the decimation process, the anti-aliasing filter which operates at a sampling rate of $L f_s$ approximates the ideal lowpass response:

$$\tilde{H}(e^{j\omega}) = \begin{cases} L, & |\omega| \leq \frac{2\pi f_s}{2L} = \frac{f_s}{L} \\ 0, & \text{otherwise} \end{cases} \quad (\text{A.3})$$

From the above description, the interpolator transfer function can then be derived as:

$$y(m) = \sum_{n=-\infty}^{\infty} h(nL + m \oplus L) x\left[\left\lfloor \frac{m}{L} \right\rfloor - n\right] \quad (\text{A.4})$$

where $\oplus$ denotes the modulo operator and $\lfloor \cdot \rfloor$ is the floor function.

Fractional rate conversion can be thought of as a cascade of the interpolation and decimation processes as shown in Fig. 2.10. It is important to note interpolation always precedes decimation such that the intermediate signal $w(k)$ has a bandwidth greater than or equal to that of either $x(n)$ or $y(m)$. In this case, the bandlimiting filter $h(k)$ which operates at a sampling rate of $L f_s$ approximates the following frequency response:

$$\tilde{H}(e^{j\omega}) = \begin{cases} L, & \left| \frac{2\pi}{L} \right| \leq \min \left| \frac{\pi}{L}, \frac{\pi}{M} \right| \\ 0, & \text{otherwise} \end{cases} \quad (\text{A.5})$$

The transfer function of the $M:L$ rate converter can be derived by combining (A.1) and (A.4):

$$y(m) = \sum_{n=-\infty}^{\infty} h(nL + mM \oplus L) x\left[\left\lfloor \frac{mM}{L} \right\rfloor - n\right] \quad (\text{A.6})$$

Needless to say, performance of the above rate converters is much influenced by the filter design. Finite Impulse Response (FIR) polyphase filters [11] which are inherently

stable and possess linear phase properties are now discussed.

A commutator model for the $1:L$ interpolator is shown in Fig. A.1(a). The FIR filter $h(n)$ in (A.4) is partitioned into L sets of polyphase filters $\{g_0(n), g_1(n), \dots, g_{L-1}(n)\}$ where:

$$g_m(n) = h(nL + m \oplus L), \quad m = 0, 1, \dots, L-1 \quad (\text{A.7})$$

For every input $x(n)$, the commutator sweeps through L branches of the commutator to obtain L output samples $y(m)$. At every branch, $x(n)$ is filtered by one of L partitioned filter sets to yield $y_m(n)$. After L output samples have been obtained, the filtering operation is repeated. Similarly, the commutator model can be obtained for the $M:1$ decimator by transposing the interpolator structure as shown in Fig. A.1(b). In this case, $h(n)$ is partitioned into M sets of polyphase filters $\{p_0(n), p_1(n), \dots, p_{M-1}(n)\}$ where:

$$p_k(n) = h(nM + k), \quad k = 0, 1, \dots, M-1 \quad (\text{A.8})$$

The commutator thus distributes M input samples of $x(m)$ in a counterclockwise manner (from $k = M-1, M-2, \dots, 1, 0$) to each branch. When all M branches have received a new input at $n = mM$, the filter outputs are summed to give $y(n)$. The same polyphase filtering is then repeated for the next M inputs.

A polyphase structure can also be derived for an $M:L$ fractional rate converter. If the FIR filter length N is a multiple of L such that $N = QL$, then (A.6) can be rewritten as:

$$y(m) = \sum_{n=0}^{Q-1} g_{m \oplus L}(n) x\left[\left\lfloor \frac{mM}{L} \right\rfloor - n\right] \quad (\text{A.9})$$

where $g_{m \oplus L}(n)$ denotes a periodic polyphase filter with Q coefficients and period L and is related to the lowpass FIR filter $h(n)$ by:

$$g_{m \oplus L}(n) = h(nL + mM \oplus L) \quad (\text{A.10})$$

(A.9) can be implemented efficiently using a sample-and-hold circuit [11] as illustrated in Fig. A.2.

In this study, equiripple FIR filters which have minimum passband ripples [11] are used to obtain the rate converters. The filter length of each of the sampling rate converters is estimated initially based on the following [15]:

(a) Polyphase Filter for 1:L Interpolator

(b) Polyphase Filter for M:1 Decimator

Figure A.1 - Polyphase Filter Networks

Figure A.2 - Modified Polyphase Filter for $M:L$ Rate Converter

At each output sample time, Q samples from the sample and hold circuits are multiplied respectively with Q coefficients from the polyphase filter and products are accumulated to obtain $y(m)$. The sample and hold circuit is updated by shifting in one new input sample whenever the floor function quantity increases by one.

$$N = \frac{-20 \log_{10} [\sqrt{\delta_1 \delta_2}] - 13}{14.6 \Delta f} + 1 \quad (\text{A.11})$$

where

$$\delta_1 = 10^{\frac{-A_{\max}}{40}}, \quad \delta_2 = 10^{\frac{-A_{\min}}{20}} \quad \text{and} \quad \Delta f = \frac{\omega_p - \omega_s}{2\pi}$$

where N = FIR filter length, $A_{\max}$ = maximum passband ripple (dB), $A_{\min}$ = minimum stopband attenuation (dB), Δf = passband range (Hz), ω_p = passband frequency (rad/s) and ω_s = stopband frequency (rad/s).

A common set of specifications for the signal BW ω_c , passband ripple $A_{\max}$ and stopband attenuation $A_{\min}$ can be chosen for the rate converters such that:

$$\omega_c = \frac{3000\pi}{4000} = 0.75\pi, \quad A_{\max} \leq 0.2\text{dB}, \quad A_{\min} \geq 35\text{dB}$$

Additionally, the normalized passband and stopband frequencies for the different sample rate conversions are tabulated below:

Sampling Rate Conversion	ω_p	ω_s
$M:L$ Rate Conversion	$\omega_p = \min\left[\frac{\omega_c}{L}, \frac{\omega_c}{M}\right]$	$\omega_s = \min\left[\frac{\pi}{L}, \frac{\pi}{M}\right]$
$1:L$ Interpolation	$\omega_p = \frac{\omega_c}{L}$	$\omega_s = \frac{\pi}{L}$
$M:1$ Decimation	$\omega_p = \frac{\omega_c}{M}$	$\omega_s = \frac{\pi}{M}$

Table A.1 - Passband and Stopband Frequencies of Sampling Rate Converters

Utilizing the Parks-McClellan equiripple FIR filter design methods [11], low pass filters of lengths $N=75$ and $N=57$ which meet the previous filter specifications were obtained for the 4:5 and 4:3 fractional rate converters. Their respective frequency responses are shown in Fig. A.4 and Fig. A.3. An identical low-pass filter design can be shared for the 1:2 interpolator and 2:1 decimator. For this purpose, a 30-tap lowpass filter with frequency response as shown in Fig. A.5 was therefore obtained.

Figure A.3 - Frequency Response of 4:3 Sample Rate Converter

Figure A.4 - Frequency Response of 4:5 Sample Rate Converter

Figure A.5 - Frequency Response of 1:2/1:2 Sample Rate Converter

Appendix B

Cholesky Decomposition

An approximation of the Wiener-Hopf Equations for an N^{th} order vector linear predictor was formulated in (4.29) as:

$$\tilde{\mathbf{R}}\mathbf{A} = -\tilde{\mathbf{V}} \quad (\text{B.1})$$

The LP coefficient matrix $\mathbf{A}$, supervector $\tilde{\mathbf{V}}$ and supermatrix $\tilde{\mathbf{R}}$ are defined as:

$$\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_N]^T \quad (\text{B.2})$$

$$\tilde{\mathbf{V}} = [\tilde{\mathbf{R}}_{10}^T, \tilde{\mathbf{R}}_{20}^T, \dots, \tilde{\mathbf{R}}_{N0}^T]^T \quad (\text{B.3})$$

$$\tilde{\mathbf{R}} = \begin{bmatrix} \tilde{\mathbf{R}}_{11} & \tilde{\mathbf{R}}_{12} & \dots & \cdot & \tilde{\mathbf{R}}_{1N} \\ \tilde{\mathbf{R}}_{21} & \tilde{\mathbf{R}}_{22} & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \cdot \\ \cdot & \cdot & \dots & \cdot & \tilde{\mathbf{R}}_{N-1N} \\ \tilde{\mathbf{R}}_{N1} & \cdot & \dots & \tilde{\mathbf{R}}_{NN-1} & \tilde{\mathbf{R}}_{NN} \end{bmatrix} \quad (\text{B.4})$$

where

$$\tilde{\mathbf{R}}_{ij} = \sum_{n=N}^M \mathbf{X}_{n-i} \mathbf{X}_{n-j}^T \quad (\text{B.5})$$

same as (4.29). The first step in solving the above equation is to reformulate (B.1) into N scalar LP problems. This could easily be done by rewriting (B.1) in terms of only the col-

umn vectors of LP matrix A and supervector V :

$$[\tilde{R}a_1, \tilde{R}a_2, \dots, \tilde{R}a_{N-1}, \tilde{R}a_N] = -[v_1, v_2, \dots, v_{N-1}, v_N] \quad (B.6)$$

where a_n and v_n for $n=1, 2, \dots, N$ are the column vectors of LP coefficient matrix A and supervector V . For a particular scalar predictor in (B.6), we need to solve:

$$\tilde{R}a_n = -v_n \quad (B.7)$$

If the supermatrix $\tilde{R}$ is positive definite, then it can be decomposed into the form:

$$\tilde{R} = LDU \quad (B.8)$$

where L is a lower triangular matrix, D is a diagonal matrix, and U is an upper triangular. Since $\tilde{R}$ is a symmetric matrix, it follows that $L = U^T$ and (B.8) can be rewritten as:

$$\begin{aligned} \tilde{R} &= (L\sqrt{D}) \cdot (\sqrt{D}U) \\ &= (L\sqrt{D}) \cdot (\sqrt{D}L^T) \\ &= W \cdot W^T \end{aligned} \quad (B.9)$$

If we substitute Equation (B.9) into (B.7):

$$\begin{aligned} WW^T(-a_n) &= v_n \\ Wy_n &= v_n \end{aligned} \quad (A.10)$$

where $y_n = W^T(-a_n)$, then we can solve for y_n , and consequently $-a_n$ via back substitution. To illustrate the procedure, a 4th order scalar LP of the form $Ra = -v$ is first written in matrix form as:

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \\ -a_3 \\ -a_4 \end{bmatrix} = \begin{bmatrix} r_{10} \\ r_{20} \\ r_{30} \\ r_{40} \end{bmatrix}$$

and $R = WW^T$ can be expressed such that:

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} \\ r_{21} & r_{22} & r_{23} & r_{24} \\ r_{31} & r_{32} & r_{33} & r_{34} \\ r_{41} & r_{42} & r_{43} & r_{44} \end{bmatrix} = \begin{bmatrix} w_{11} & 0 & 0 & 0 \\ w_{21} & w_{22} & 0 & 0 \\ w_{31} & w_{32} & w_{33} & 0 \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix} \begin{bmatrix} w_{11} & w_{21} & w_{31} & w_{41} \\ 0 & w_{22} & w_{32} & w_{42} \\ 0 & 0 & w_{33} & w_{43} \\ 0 & 0 & 0 & w_{44} \end{bmatrix}$$

and (B.10) becomes:

$$\begin{bmatrix} w_{11} & 0 & 0 & 0 \\ w_{21} & w_{22} & 0 & 0 \\ w_{31} & w_{32} & w_{33} & 0 \\ w_{41} & w_{42} & w_{43} & w_{44} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} r_{10} \\ r_{20} \\ r_{30} \\ r_{40} \end{bmatrix}$$

After y_n is found via back substitution, $-a_n$ can be obtained as:

$$\begin{bmatrix} w_{11} & w_{21} & w_{31} & w_{41} \\ 0 & w_{22} & w_{32} & w_{42} \\ 0 & 0 & w_{33} & w_{43} \\ 0 & 0 & 0 & w_{44} \end{bmatrix} \begin{bmatrix} -a_1 \\ -a_2 \\ -a_3 \\ -a_4 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

From the above, it can be deduced that r_{ij} can be written as:

$$r_{ij} = \sum_{k=1}^j w_{ik} w_{jk} = w_{ij} w_{jj} + \sum_{k=1}^{j-1} w_{ik} w_{jk}, \quad 1 \leq j \leq i \quad (\text{B.11})$$

and w_{ij} can be derived as:

$$w_{ij} = \frac{\left[r_{ij} - \sum_{k=1}^{j-1} w_{ik} w_{jk} \right]}{w_{jj}}, \quad 1 \leq j < i \quad (\text{B.12})$$

where w_{jj} is obtained from (B.11) as:

$$w_{jj} = \sqrt{r_{jj} - \sum_{k=1}^{j-1} w_{jk}^2} \quad (\text{B.13})$$

The strategy is therefore to compute the lower triangular matrix $\mathbf{W}$ in columns from top to bottom. Once $\mathbf{W}$ is found, the column vector $\mathbf{y}$ and hence the LP coefficient vector $\mathbf{a}$ are given respectively by:

$$y_i = \left[r_{i0} - \sum_{j=1}^{i-1} w_{ij} y_j \right] / w_{ii}, \quad i = 1, 2, \dots, p \quad (\text{B.14})$$

$$a_i = - \left[y_i - \sum_{j=i+1}^p w_{ji} a_j \right] / w_{ii}, \quad i = p, p-1, \dots, 1 \quad (\text{B.15})$$

where p is the dimension of the column vectors $\mathbf{y}$ and $\mathbf{a}$. It can be noted that y_i is computed in ascending order whereas a_i is computed in descending order. For the N^{th} order VLP, the above procedures are repeated for all $N - a_n$ columns in (B.6) and a solution for $\mathbf{A}$ can be obtained.

As a final note, the coefficient vector $\mathbf{a}$ may generate unstable linear predictor. A modified Cholesky procedure may be used to ensure stability by transforming y_i into the i^{th} reflection coefficient [34]:

$$k_i = \frac{y_i}{\sqrt{D_i}}, \quad i = 1, 2, \dots, p \quad (\text{B.16})$$

where D_i is the squared prediction error at the i^{th} step of the prediction and is given by:

$$\begin{aligned} D_i &= r_{00} - \mathbf{a}^T \mathbf{R} \mathbf{a} = r_{00} - \mathbf{a}^T \mathbf{W} \mathbf{W}^T \mathbf{a} \\ &= r_{00} - \mathbf{y} \mathbf{y}^T = r_{00} - \sum_{i=1}^p y_i^2 \end{aligned} \quad (\text{B.17})$$

Since $|k_i| \leq 1$, the resulting predictor is stable. The p reflection coefficients found can be transformed back into a_i easily (- see Section 4.2.3.) It should be noted the resulting predictor is non-optimum in minimizing the mean squared prediction error.

Appendix C

Levinson-Wiggins-Robinson (LWR) Recursion

The LWR algorithm which was developed for multi-channel Wiener filtering of geophysical information is an efficient method for joint-process estimation. In the original application, data signals are received from multiple channels for example, an array of seismographs and are used to provide least-squares fits for some desired outputs. In the N^{th} order VLP problem, past vector samples are used to estimate the desired output which is the current vector sample.

The following two-step derivation is based on the paper by Wiggins and Robinson [63][4]. In the first step, an auxiliary polynomial recursion (similar to Levinson-Durbin recursion) is established and the second step makes use of the auxiliary coefficients to provide the multi-channel filter recursion. First of all, the VLP formulation in (4.28) is written in matrix form using the autocorrelation matrix expression of (4.30):

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{N-1} \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_{N-1} & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_{N-1} \end{bmatrix} \begin{bmatrix} \mathbf{A}_1^{(N)T} \\ \mathbf{A}_2^{(N)T} \\ \cdot \\ \cdot \\ \mathbf{A}_N^{(N)T} \end{bmatrix} = - \begin{bmatrix} \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_2 \\ \cdot \\ \cdot \\ \tilde{\mathbf{R}}_N \end{bmatrix} \quad (\text{C.1})$$

where

$$\tilde{\mathbf{R}}_i = \sum_{n=1+i}^M \mathbf{X}_n \mathbf{X}_{n-i}^T \quad (\text{C.2})$$

The superscript (N) in (C.1) is used to denote $\mathbf{A}_i$ is the i^{th} LP coefficient matrix of an N^{th} order VLP.

From the theory of autoregressive modelling [4], the forward and backward N^{th} order prediction error matrices α_n and β_n for the vector process $\{\mathbf{X}_n\}$ can be defined respectively as:

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_N \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_N & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{a}_{0n} \\ \mathbf{a}_{1n} \\ \cdot \\ \cdot \\ \mathbf{a}_{nn} \end{bmatrix} = \begin{bmatrix} \alpha_n \\ \mathbf{0} \\ \cdot \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix} \quad (\text{C.3})$$

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_N \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_N & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{b}_{nn} \\ \cdot \\ \cdot \\ \mathbf{b}_{1n} \\ \mathbf{b}_{0n} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \cdot \\ \mathbf{0} \\ \beta_n \end{bmatrix} \quad (\text{C.4})$$

where $\mathbf{a}_{in}$ and $\mathbf{b}_{in}$ are the auxiliary coefficient matrices with $\mathbf{a}_{0n} = \mathbf{b}_{0n} = \mathbf{I}$ and $\mathbf{0}$ is the zero matrix.

Extending the prediction errors to the $(N+1)^{\text{st}}$ case leads to the results:

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{N+1} \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_{N+1} & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{a}_{0n} \\ \mathbf{a}_{1n} \\ \cdot \\ \mathbf{a}_{nn} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \alpha_n \\ \mathbf{0} \\ \cdot \\ \mathbf{0} \\ \alpha'_n \end{bmatrix} \quad (\text{C.5})$$

and

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{N+1} \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_{N+1} & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{b}_{nn} \\ \vdots \\ \mathbf{b}_{1n} \\ \mathbf{b}_{0n} \end{bmatrix} = \begin{bmatrix} \beta'_n \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \beta_n \end{bmatrix} \quad (\text{C.6})$$

(C.5) and (C.6) can be combined linearly using scaling matrices $\mathbf{K}_{an}$ and $\mathbf{K}_{bn}$ (note: these are equivalent to reflection coefficients) to yield the following pair of equations:

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{N+1} \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_{N+1} & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{a}_{0n} \\ \mathbf{a}_{1n} + \mathbf{K}_{an} \mathbf{b}_{nn} \\ \cdot \\ \mathbf{a}_{nn} + \mathbf{K}_{an} \mathbf{b}_{1n} \\ \mathbf{K}_{an} \mathbf{b}_{0n} \end{bmatrix} = \begin{bmatrix} \alpha'_n + \mathbf{K}_{an} \beta'_n \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \alpha'_n + \mathbf{K}_{an} \beta_n \end{bmatrix} \quad (\text{C.7})$$

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_{N+1} \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_{N+1} & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{K}_{bn} \mathbf{a}_{0n} \\ \mathbf{b}_{nn} + \mathbf{K}_{bn} \mathbf{a}_{1n} \\ \cdot \\ \mathbf{b}_{1n} + \mathbf{K}_{bn} \mathbf{a}_{nn} \\ \mathbf{b}_{0n} \end{bmatrix} = \begin{bmatrix} \beta'_n + \mathbf{K}_{bn} \alpha'_n \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \beta_n + \mathbf{K}_{bn} \alpha'_n \end{bmatrix} \quad (\text{C.8})$$

In order to minimize the $(N+1)^{\text{st}}$ prediction errors, $\mathbf{K}_{an}$ and $\mathbf{K}_{bn}$ must be chosen such that:

$$\alpha'_n + \mathbf{K}_{an} \beta_n = \mathbf{0} \quad \beta_n + \mathbf{K}_{bn} \alpha'_n = \mathbf{0} \quad (\text{C.9a,b})$$

Subsequently, the auxiliary LP filter coefficient matrices are given by:

$$\begin{aligned} & [\mathbf{a}_{0,n+1}, \mathbf{a}_{1,n+1}, \dots, \mathbf{a}_{n+1,n+1}] \\ &= [\mathbf{a}_{0,n}, \dots, \mathbf{a}_{n,n}, \mathbf{0}] + \mathbf{K}_{an} [\mathbf{0}, \mathbf{b}_{nn}, \dots, \mathbf{b}_{0n}] \end{aligned} \quad (\text{C.10})$$

$$\begin{aligned} & [\mathbf{b}_{n+1,n+1}, \dots, \mathbf{b}_{1,n+1}, \mathbf{b}_{0,n+1}] \\ &= [\mathbf{0}, \mathbf{b}_{n,n}, \dots, \mathbf{b}_{0,n}] + \mathbf{K}_{bn} [\mathbf{a}_{0,n}, \dots, \mathbf{a}_{n,n}, \mathbf{0}] \end{aligned} \quad (\text{C.11})$$

From the above equations, the $(N+1)^{\text{st}}$ LP auxiliary coefficient matrices are:

$$\begin{aligned} \mathbf{a}_{n+1, n+1} &= -\alpha'_n \beta_n^{-1} = \mathbf{K}_{an} \\ \mathbf{b}_{n+1, n+1} &= -\beta'_n \alpha_n^{-1} = \mathbf{K}_{bn} \end{aligned} \quad (\text{C.12a,b})$$

Additionally, the $(N+1)^{\text{st}}$ prediction error matrices α_{n+1} and β_{n+1} can be extracted from (C.7) and (C.8) and they are given by:

$$\begin{aligned} \alpha_{n+1} &= \alpha_n + \mathbf{K}_{an} \beta'_n \\ \beta_{n+1} &= \beta_n + \mathbf{K}_{bn} \alpha'_n \end{aligned} \quad (\text{C.13 a,b})$$

The first step of the recursion is hereby completed and the auxiliary coefficients $\{\mathbf{b}_{i,n}\}$ can be used to establish the general filter recursion in a similar fashion. The Wiener-Hopf matrix equation in (C.1) is therefore extended to $(N+1)^{\text{st}}$ order:

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_N \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_N & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{A}_1^{(N)T} \\ \mathbf{A}_2^{(N)T} \\ \cdot \\ \mathbf{A}_N^{(N)T} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} -\tilde{\mathbf{R}}_1 \\ -\tilde{\mathbf{R}}_2 \\ \cdot \\ -\tilde{\mathbf{R}}_N \\ \gamma_N \end{bmatrix} \quad (\text{C.14})$$

A linear combination of $\mathbf{A}^{(N)}$ and $\mathbf{b}_n$ then gives:

$$\begin{bmatrix} \tilde{\mathbf{R}}_0 & \tilde{\mathbf{R}}_1 & \dots & \tilde{\mathbf{R}}_N \\ \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \tilde{\mathbf{R}}_1 \\ \tilde{\mathbf{R}}_N & \cdot & \dots & \tilde{\mathbf{R}}_1 & \tilde{\mathbf{R}}_0 \end{bmatrix} \begin{bmatrix} \mathbf{A}_1^{(N)T} + \mathbf{K}_{fn} \mathbf{b}_{n,n} \\ \mathbf{A}_2^{(N)T} + \mathbf{K}_{fn} \mathbf{b}_{n-1,n} \\ \cdot \\ \mathbf{A}_N^{(N)T} + \mathbf{K}_{fn} \mathbf{b}_{1,n} \\ \mathbf{K}_{fn} \mathbf{b}_{0,n} \end{bmatrix} = \begin{bmatrix} -\tilde{\mathbf{R}}_1 \\ -\tilde{\mathbf{R}}_2 \\ \cdot \\ -\tilde{\mathbf{R}}_N \\ \gamma_N + \mathbf{K}_{fn} \beta_n \end{bmatrix} \quad (\text{C.15})$$

The $(N+1)^{\text{st}}$ order VLP is obtained by setting:

$$\gamma_N + \mathbf{K}_{fn} \beta_n = -\tilde{\mathbf{R}}_{N+1} \quad (\text{C.16})$$

and the new VLP coefficients can be found to be:

$$\begin{aligned} & [\mathbf{A}_1^{(N+1)T}, \dots, \mathbf{A}_N^{(N+1)T}, \mathbf{A}_{N+1}^{(N+1)T}] \\ &= [\mathbf{A}_1^{(N)T}, \dots, \mathbf{A}_N^{(N)T}, \mathbf{0}] + \mathbf{K}_{/n} [\mathbf{b}_{n,n}, \dots, \mathbf{b}_{1,n}, \mathbf{0}] \end{aligned} \quad (\text{C.17})$$

The above thus provides a way to compute the LP coefficient matrices recursively from VLP of lower order. Like the Levinson-Durbin algorithm, the above derivation also yields an expression for an error matrix at every stage of the prediction. Denoting the error matrix as $\mathbf{E}_N$, it can be shown that:

$$\mathbf{E}_N = \tilde{\mathbf{R}}_0 - (\mathbf{A}_1^{(N)} \tilde{\mathbf{R}}_1 + \mathbf{A}_2^{(N)} \tilde{\mathbf{R}}_2 + \dots + \mathbf{A}_N^{(N)} \tilde{\mathbf{R}}_N) \quad (\text{C.18})$$

Appendix D

Mean Opinion Score (MOS) Test Instructions

Dear Friends and Colleagues:

Thank you for participating in this listening test session.

These tests are designed to examine the effects of a variety of telephone transmission impairments on speech quality. A number of American-English recordings spoken by different talkers will be played. Your task is to listen to these test conditions and rate how good the speech sounds as if it were a normal telephone conversation (NOTE: only left-side of the headphone is used.)

The MOS test procedures are listed as follows:

- **The MOS test consists of 8 trials (including one practice run) each with 13 test conditions. The practice is to let you become familiarized with the range of speech quality you will encounter in the test.**
- **At the beginning of a test condition, a beep will sound and this will be followed by 10 secs. of speech and a second beep. You should score by *circling* the appropriate number on the given form during the silence interval in between the end of recording and the second beep. Two consecutive beeps denote the end of current trial.**
- **Your score should be based on your total impression of each recording, rather than**

concentrating on particular aspects. The speech quality should be rated according to the scale below:

- You should *never* refer to your previous scores when you are scoring for the current test condition.
- If you find you have missed scoring one of the test conditions, simply stop further scoring until the current test trial is completed
- There is a pause in between each trial to let you ask the instructor questions regarding your test progress.

END

21-07-94

FIN