Code and Wavelet Excited Linear Prediction of Speech

by Brendan Frey

A thesis submitted to the Faculty of Graduate Studies
in partial fulfillment of the thesis requirements
for the degree of

Master of Science
in
Electrical Engineering

Department of Electrical and Computer Engineering
University of Manitoba
Winnipeg, Canada 1993

ISBN 0-315-86017-0

Name ______________________________

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.

Electronics & Electrical Engineering.
SUBJECT TERM

0554
SUBJECT CODE

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

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	0984
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	0554
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

Nom ____

Dissertation Abstracts International est organisé en catégories de sujets. Veuillez s.v.p. choisir le sujet qui décrit le mieux votre thèse et inscrivez le code numérique approprié dans l'espace réservé ci-dessous.

____ □□□□ U·M·I

SUJET CODE DE SUJET

Catégories par sujets

HUMANITÉS ET SCIENCES SOCIALES

COMMUNICATIONS ET LES ARTS

Architecture	0729
Beaux-arts	0357
Bibliothéconomie	0399
Cinéma	0900
Communication verbale	0459
Communications	0708
Danse	0378
Histoire de l'art	0377
Journalisme	0391
Musique	0413
Sciences de l'information	0723
Théâtre	0465

ÉDUCATION

Généralités	515
Administration	0514
Art	0273
Collèges communautaires	0275
Commerce	0688
Économie domestique	0278
Éducation permanente	0516
Éducation préscolaire	0518
Éducation sanitaire	0680
Enseignement agricole	0517
Enseignement bilingue et multiculturel	0282
Enseignement industriel	0521
Enseignement primaire.	0524
Enseignement professionnel	0747
Enseignement religieux	0527
Enseignement secondaire	0533
Enseignement spécial	0529
Enseignement supérieur	0745
Évaluation	0288
Finances	0277
Formation des enseignants	0530
Histoire de l'éducation	0520
Langues et littérature	0279
Lecture	0535
Mathématiques	0280
Musique	0522
Orientation et consultation	0519
Philosophie de l'éducation	0998
Physique	0523
Programmes d'études et enseignement	0727
Psychologie	0525
Sciences	0714
Sciences sociales	0534
Sociologie de l'éducation	0340
Technologie	0710

LANGUE, LITTÉRATURE ET LINGUISTIQUE

Langues	
Généralités	0679
Anciennes	0289
Linguistique	0290
Modernes	0291
Littérature	
Généralités	0401
Anciennes	0294
Comparée	0295
Mediévale	0297
Moderne	0298
Africaine	0316
Américaine	0591
Anglaise	0593
Asiatique	0305
Canadienne (Anglaise)	0352
Canadienne (Française)	0355
Germanique	0311
Latino-américaine	0312
Moyen-orientale	0315
Romane	0313
Slave et est-européenne	0314

PHILOSOPHIE, RELIGION ET THÉOLOGIE

Philosophie	0422
Religion	
Généralités	0318
Clergé	0319
Études bibliques	0321
Histoire des religions	0320
Philosophie de la religion	0322
Théologie	0469

SCIENCES SOCIALES

Anthropologie	
Archéologie	0324
Culturelle	0326
Physique	0327
Droit	0398
Économie	
Généralités	0501
Commerce-Affaires	0505
Économie agricole	0503
Économie du travail	0510
Finances	0508
Histoire	0509
Théorie	0511
Études américaines	0323
Études canadiennes	0385
Études féministes	0453
Folklore	0358
Géographie	0366
Gérontologie	0351
Gestion des affaires	
Généralités	0310
Administration	0454
Banques	0770
Comptabilité	0272
Marketing	0338
Histoire	
Histoire générale	0578
Ancienne	0579
Médiévale	0581
Moderne	0582
Histoire des noirs	0328
Africaine	0331
Canadienne	0334
États-Unis	0337
Européenne	0335
Moyen-orientale	0333
Latino-américaine	0336
Asie, Australie et Océanie	0332
Histoire des sciences	0585
Loisirs	0814
Planification urbaine et régionale	0999
Science politique	
Généralités	0615
Administration publique	0617
Droit et relations internationales	0616
Sociologie	
Généralités	0626
Aide et bien-àtre social	0630
Criminologie et établissements pénitentiaires	0627
Démographie	0938
Études de l' individu et de la famille	0628
Études des relations interethniques et des relations raciales	0631
Structure et développement social	0700
Théorie et méthodes.	0344
Travail et relations industrielles	0629
Transports	0709
Travail social	0452

SCIENCES ET INGÉNIERIE

SCIENCES BIOLOGIQUES

Agriculture	
Généralités	0473
Agronomie.	0285
Alimentation et technologie alimentaire	0359
Culture	0479
Élevage et alimentation	0475
Exploitation des péturages	0777
Pathologie animale	0476
Pathologie végétale	0480
Physiologie végétale	0817
Sylviculture et faune	0478
Technologie du bois	0746
Biologie	
Généralités	0306
Anatomie	0287
Biologie (Statistiques)	0308
Biologie moléculaire	0307
Botanique	0309
Cellule	0379
Écologie	0329
Entomologie	0353
Génétique	0369
Limnologie	0793
Microbiologie	0410
Neurologie	0317
Océanographie	0416
Physiologie	0433
Radiation	0821
Science vétérinaire	0778
Zoologie	0472
Biophysique	
Généralités	0786
Medicale	0760

SCIENCES DE LA TERRE

Biogéochimie	0425
Géochimie	0996
Géodésie	0370
Géographie physique	0368
Géologie	0372
Géophysique	0373
Hydrologie	0388
Minéralogie	0411
Océanographie physique	0415
Paléobotanique	0345
Paléoécologie	0426
Paléontologie	0418
Paléozoologie	0985
Palynologie	0427

SCIENCES DE LA SANTÉ ET DE L'ENVIRONNEMENT

Économie domestique	0386
Sciences de l'environnement	0768
Sciences de la santé	
Généralités	0566
Administration des hipitaux	0769
Alimentation et nutrition	0570
Audiologie	0300
Chimiothérapie	0992
Dentisterie	0567
Développement humain	0758
Enseignement	0350
Immunologie	0982
Loisirs	0575
Médecine du travail et thérapie	0354
Médecine et chirurgie	0564
Obstétrique et gynécologie	0380
Ophtalmologie	0381
Orthophonie	0460
Pathologie	0571
Pharmacie	0572
Pharmacologie	0419
Physiothérapie	0382
Radiologie	0574
Santé mentale	0347
Santé publique	0573
Soins infirmiers	0569
Toxicologie	0383

SCIENCES PHYSIQUES

Sciences Pures

Chimie	
Genéralités	0485
Biochimie	487
Chimie agricole	0749
Chimie analytique	0486
Chimie minérale	0488
Chimie nucléaire	0738
Chimie organique	0490
Chimie pharmaceutique	0491
Physique	0494
PolymÇres	0495
Radiation	0754
Mathématiques	0405
Physique	
Généralités	0605
Acoustique	0986
Astronomie et astrophysique	0606
Electronique et électricité	0607
Fluides et plasma	0759
Météorologie	0608
Optique	0752
Particules (Physique nucléaire)	0798
Physique atomique	0748
Physique de l'état solide	0611
Physique moléculaire	0609
Physique nucléaire	0610
Radiation	0756
Statistiques	0463

Sciences Appliqués Et Technologie

Informatique	0984
Ingénierie	
Généralités	0537
Agricole	0539
Automobile	0540
Biomédicale	0541
Chaleur et ther modynamique	0348
Conditionnement (Emballage)	0549
Génie aérospatial	0538
Génie chimique	0542
Génie civil	0543
Génie électronique et électrique	0544
Génie industriel	0546
Génie mécanique	0548
Génie nucléaire	0552
Ingénierie des systämes	0790
Mécanique navale	0547
Métallurgie	0743
Science des matériaux	0794
Technique du pétrole	0765
Technique minière	0551
Techniques sanitaires et municipales	0554
Technologie hydraulique	0545
Mécanique appliquée	0346
Géotechnologie	0428
Matières plastiques (Technologie)	0795
Recherche opérationnelle	0796
Textiles et tissus (Technologie)	0794

PSYCHOLOGIE

Généralités	0621
Personnalité	0625
Psychobiologie	0349
Psychologie clinique	0622
Psychologie du comportement	0384
Psychologie du développement	0620
Psychologie expérimentale	0623
Psychologie industrielle	0624
Psychologie physiologique	0989
Psychologie sociale	0451
Psychométrie	0632

CODE AND WAVELET EXCITED LINEAR PREDICTION OF SPEECH

BY

BRENDAN FREY

A Thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

I hereby declare that I am the sole author of this thesis.

I authorize the University of Manitoba to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I also authorize the University of Manitoba to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Brendan Frey, 1993.

Abstract

Speech compression for the purposes of storage and transmission has long been important, especially in the telecommunications industry. Coding techniques range from the computationally simple (μ-law, ADPCM) to the computationally complex (CELP). Most sophisticated techniques use linear prediction to obtain a model of the speaker's vocal tract for each of a series of 20 ms (or so) speech segments.

For the method of linear predictive coding (LPC), it is assumed that either a periodic impulse train or white noise is sufficient to model the glottal excitation. This method requires a DSP chip rating of about 1.7 MIPS to provide reasonable speech quality (somewhat mechanical sounding) at 2400 bps.

Code Excited Linear Prediction (CELP) is similar to LPC, but it accounts for nonideal excitation. For each code in a codebook, the CELP compressor applies the vocal tract filter, subtracts the result from the real speech to obtain an error signal, perceptually weights the error signal, and then calculates the norm of the weighted error. The code with the minimum weighted error is considered optimal and its index and gain are transmitted over the channel. This method requires a DSP chip rating of about 25 MIPS to provide excellent speech quality at 4800 bps. Although high-MIPS DSPs are available for CELP implementations, their power requirements would heavily burden the small battery that would be available in a cellular phone or a portable terminal. A lower-power system is better suited to the portable communication link.

Presented in this thesis is an original compression scheme for which the error feedback loop is closed earlier than for CELP. Instead of determining the best approximation to the speech segment, the best approximation to the vocal tract excitation is found. As well, in order to reduce the codebook search time, error not accounted for by a simple long-term codebook is coded using the discrete wavelet transform (DWT). This method is called Code and Wavelet Excited Linear Prediction (CWELP). It was motivated by intentions to improve the speech quality of the LPC vocoder while using less computation than the CELP vocoder. The CWELP vocoder requires 3.6 MIPS and provides speech quality slightly superior to LPC at a bit rate of 4600 bps.

Individual letter recognition tests were performed using the ISOLET database. Whereas the recognition accuracy for LPC and CELP was 73% and 85%, the accuracy for CWELP was 76%. Informal subjective sentence tests show that the CWELP-compressed speech sounds slightly better than the LPC-compressed speech. Tests show that because it models more complicated forms of excitation, CWELP can compress non-human sounds better than LPC can. CWELP appears to be in its expected performance regime: a compromise between LPC and CELP.

Acknowledgments

I appreciate the guidance from Dr. Howard Card; the assistance from SriGouri Kamarsu, Ken Ferens, Dean McNeill and Alex McIlraith; and the ears lent to me by part of the electrical engineering department at the University of Calgary for hearing a dry run of my thesis presentation.

I also greatly value the financial support by NSERC.

Contents

Abstract iii

Acknowledgments iv

List of Figures vii

List of Tables ix

Chapter 1: Introduction 1

Chapter 2: Physiology of Vocalization 3

- 2.1 Structure of the Vocal System 3
- 2.2 Phonemes 4
- 2.3 Analysis of the Vocal System 6
  - 2.3.1 Voiced vs Unvoiced Speech 6
  - 2.3.2 Vocal Tract 7

Chapter 3: Frequency-Time Representation of Sounds 11

- 3.1 What Does Frequency-Time Mean? 11
- 3.2 Suitability of Frequency-Time Analysis to the Human Speech Channel 12
- 3.3 Sound Spectrograph 12
- 3.4 Short-Time Fourier Transform (STFT) 13
  - 3.4.1 STFT Does NOT Have an Orthonormal Basis 15
  - 3.4.2 Phase Space Localization of Basis States 16
- 3.5 Wavelet Transform 17
  - 3.5.1 Generation of the Basis from the Mother Wavelet 17
  - 3.5.2 A Simple Example: the Haar Wavelet 18
  - 3.5.3 Daubechies Discrete Wavelet Transform (DWT) 19

Chapter 4: Representation of Speech 23

- 4.1 Optimal PCM 23
- 4.2 Wavelet Representation 25
- 4.3 Wavelet Transform Compressed 25
- 4.4 Linear Predictive Coding (LPC) 26
- 4.5 Code Excited Linear Prediction (CELP) 28
  - 4.5.1 True Excitation: More Than White Noise or Impulses 28
  - 4.5.2 Using a Code to Approximate the True Excitation 29
  - 4.5.3 Important: Pre VS Post Filter Error 29
  - 4.5.4 Analysis by Synthesis CELP 30

Chapter 5: Code and Wavelet Excited Linear Prediction (CWELP) 33

5.1 Introduction 33
5.2 CWELP System 33
5.2.1 Qualitative Difference Between Short-term and Long-term Codes 35
5.3 Long-term Codebook 36
5.3.1 Coding the Code Index and Gain 37
5.4 Short-term Excitation 40
5.4.1 Coding the Short-term Excitation 41
5.5 Discussion of Synthesized Excitation 42
5.6 Bit Rate 43
5.7 Computational Complexity 44

Chapter 6: Performance 45

6.1 Comparison of Bit Rates 45
6.2 Comparison of Computational Complexity 45
6.3 Recognition Tests Using the ISOLET Database 46
6.3.1 Common Mistakes 47
6.4 Informal Sentence Tests 47
6.5 Non-human Speech Coding 47

Chapter 7: Conclusions 49

Appendix A: List of English Phonemes 51

Appendix B: Using Daubechies' Wavelet Filter to Generate the DWT 53

Appendix C: Mathematics of Classical Linear Prediction 57

Appendix D: The ISOLET Spoken Letter Database 59

References 61

List of Figures

Figure 1. Structure of the human vocal system 4

Figure 2. Waveforms for 128 ms segments from the words “hat”, “beet”, and “show”. 5

Figure 3. FFT of the phoneme // (hat). The peaks in the FFT correspond to the resonances of the vocal tract and the frequencies are called “formants”. 6

Figure 4. A typical vocal tract area function (smooth curve) and a concatenated-tube approximation. 7

Figure 5. Forces due to pressure act on a disk of air inside the vocal tract. 8

Figure 6. Electrical analogy of the vocal tract for a nasal sound. Zn is the free-air interface impedance at the nose. Note that the line impedances vary with space. 10

Figure 7. Sound spectrograph spectrogram of the utterances /UH-F-A/, /UH-S-A/, and /UH-SH-A/ from Rabiner and Schafer [1]. 13

Figure 8. Block diagram of the STFT operation. Each signal sample is multiplied by a frequency-domain basis function as with the Fourier series. Then, a filter with a window-type impulse response is used to sum a portion of the weighted sample sequence. In this way, the frequency content of the signal in the locality of each sample is calculated. 14

Figure 9. Computer-generated STFT sound spectrogram from Oppenheim [6]. It closely resembles the spectrograph output. 14

Figure 10. Location of STFT basis states (a) and wavelet transform basis states (b) in phase space (frequency-time space). 16

Figure 11. The wavelet basis is formed by shifting and then stretching or squeezing the mother wavelet. The scale for shifting is determined by the stretching or squeezing, so that high-frequency wavelets are shifted more finely. The values for m,n are shown beside each wavelet. 18

Figure 12. Time-domain basis states for the DWT generated from Daubechies’ 20-coefficient wavelet filter. 20

Figure 13. The FFT of the Haar basis states (a) indicates a greater amount of overlap in the frequency domain than for the DAUB20 basis states (b). 20

Figure 14. Time-domain basis states for the DWT generated from Daubechies’ 4-coefficient wavelet filter. 21

Figure 15. The DWT of // (hat). Notice the finer temporal resolution at higher frequencies. 21

Figure 16. Code utilization for 8-bit uniform PCM. This curve shows the probability distribution for the uniformly assigned codes. 24

Figure 17. Block diagram of the LPC synthesizer. 26

Figure 18. Comparison between the FFT of an actual phoneme, // (hat) and the FFT of its LPC synthesized counterpart. The formants are preserved. 27

Figure 19. DWTs of the original utterance of /p/ (a) and the LPC synthesized utterance (b). The frame loses the short-time features of the stop consonant. 28

Figure 20. By applying the inverse vocal tract filter to the phoneme //, we see that the true excitation is more complex than a series of impulses. 29

Figure 21. Block diagram of the CELP Analyzer. 30

Figure 22. Block diagram of a CELP synthesizer. 30

Figure 23. Block diagram of the CWELP synthesizer. 34

Figure 24. Block diagram of the CWELP analyzer. 34

Figure 25. Projecting the excitation onto the code. 36

Figure 26. True excitation (a) and long-term codebook excitation (b). With time, the codebook stores useful information about the excitation. 37

Figure 27. Probability distribution of the code index, i. 38

Figure 28. Probability distributions of the absolute code gain (a) and the code gain distribution model (b), where x is the absolute code gain. 38

Figure 29. Symbol utilization for the uniform quantization scheme (a) and the non-uniform scheme (b). 39

Figure 30. Short-term excitation (true excitation - long-term codebook excitation). 40

Figure 31. Wavelet transform of the short-term excitation. The excitation (Figure 30) consists of many short-time spikes and it is well-suited to compression in the wavelet domain. 40

Figure 32. Amplitude probability distributions of the second-largest (a) and the five next largest (b) wavelet coefficients, where x is the coefficient value. 41

Figure 33. Short-term excitation (a) and the coded short-term excitation (b). 42

Figure 34. Comparison of the true excitation (a) and the synthesized excitation (b). 43

Figure 35. Qualitative comparison between LPC, CWELP, and CELP. "Resources" refer to computational complexity and bit rate. 49

List of Tables

Table 1. Summary of bits transmitted per CWELP frame 43

Table 2. Theoretical Computational complexity of the CWELP analyzer 44

Table 3. Comparison of bit rates 45

Table 4. Comparison of computational complexity 45

Table 5. Comparison of recognition rates 46

Table 6. Common misidentifications 47

Table 7. Classification of English phonemes along with examples 51

Table 8. List of speakers used in this thesis 59

Chapter 1: Introduction

In the past, public communications have been coded using simple schemes such as A-law and μ-law. These schemes require very little computation and can be implemented cheaply in hardware. More sophisticated compression methods have been used by the U.S. government, notably linear predictive techniques used by their LPC-10e vocoder (2400 bps). When classical linear prediction (Appendix C) is applied to a speech segment, it turns out that the predictor models the vocal tract transfer function and the prediction error is the excitation for the vocal tract. The linear predictive coding (LPC) technique presumes a simple form for this excitation: white noise or a periodic impulse train. This method is consequently not very robust when presented with noisy input, multiple speakers, and non-voiced sounds.

Opening up of the telecommunications market and new high-speed DSP processors have stimulated interest in sophisticated compression schemes for use in the public network. A most impressive compression competitor is CELP which uses an exhaustive codebook search in an attempt to find a good approximation to the vocal tract excitation. The vocal tract model parameters and the codebook data are used to reconstruct the speech, segment by segment. A typical CELP system requires a 25 MIP DSP chip. This method is computational intensive because for each code the speech segment must be synthesized and perceptually weighted to determine its error measure.

Although high-MIPS DSPs are available, they are relatively expensive and have high power consumption rates. A compromise between the extremes of low speech quality, computationally simple LPC and high speech quality, computationally complex CELP seems in order. A wide variety of coding schemes have been developed and it would be futile to attempt to list them here. They range from vector quantization of blocks of LPC data frames to binary pulse CELP and so on.

The compromise presented in this thesis is Code and Wavelet Excited Linear Prediction (CWELP). Instead of determining the code which produces the "optimal" speech segment, CWELP determines the code that best matches the real vocal tract excitation. Although this leads to lower speech quality, the CWELP vocoder does not need to synthesize the speech for each code in the codebook. This shortcut reduces the computational complexity of the system significantly. The remaining excitation error is coded by transmitting the locations and magnitudes of the two largest coefficients of the excitation error's discrete wavelet transform.

Chapter 2: Physiology of Vocalization

In order to design optimal speech compression algorithms, we must understand the processes of vocalization and audition. This chapter examines the physiology of vocalization; the field of audition is not considered in this thesis although it does play an important part in many new speech compression schemes. In the general case where all signals band-limited by a particular frequency f_0 are to be coded, a sampling frequency of $f_S = 2f_0$ (Nyquist rate) is necessary and the sampled signal will use the channel to its fullest (no compression will be possible). However, by studying the physiology of speech and audition, it is possible to achieve further compression due to the physiological constraints of the animal.

2.1 Structure of the Vocal System

Figure 1 shows the physiology of the human vocal system. The lungs force air up the trachea, through the glottis and the vocal tract, and out through the lips or the nose. The glottis is the source of excitation for the vocal tract. Two types of excitation are possible giving rise to voiced and unvoiced speech. For voiced sounds, the glottis vibrates according to the tension in the vocal chords, producing a periodic signal (Figure 2a). For unvoiced sounds, the glottis is held open and the air travels along the vocal tract until it encounters a constriction (at the back of the mouth, for /sh/ as in "show" illustrated in Figure 2c). At this constriction broadband noise is generated which excites the vocal tract.

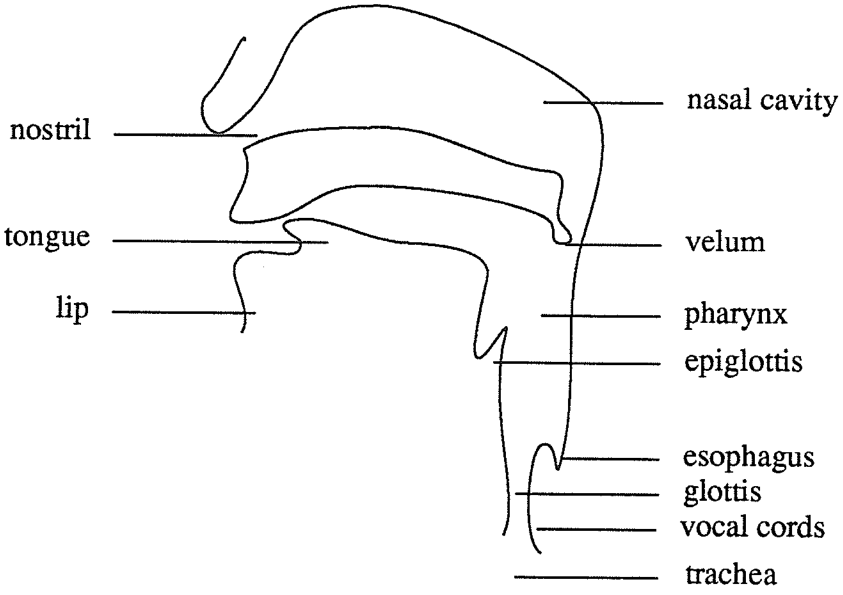


FIGURE 1. Structure of the human vocal system

By varying the shape of the vocal tract, the excitation is transformed to produce the final sound. Nasal sounds are obtained by opening the velum and closing the mouth tract (for example with the tongue, as in /n/). Finally, the sound is radiated at the lips or at the nose for nasal sounds.

2.2 Phonemes

A set of phonemes forms the atomic basis for a language. By concatenating phonemes, any utterance from the language can be constructed. Different languages often have extremely different phonetic structure. A table of the English phonemes is illustrated in Appendix A. This set of phonemes can be broken down into vowels, diphthongs, semivowels, and consonants.

Vowels are generated by fixing the shape of the vocal tract and exciting it with a periodic glottal signal. Examples are /æ/ (hat) and /i/ (beet). The periodicity of the resulting sounds is evident in Figures 2a and 2b respectively. Notice that the vowels shown have approximately the same pitch period. They sound different because each is spoken with a different vocal tract shape.

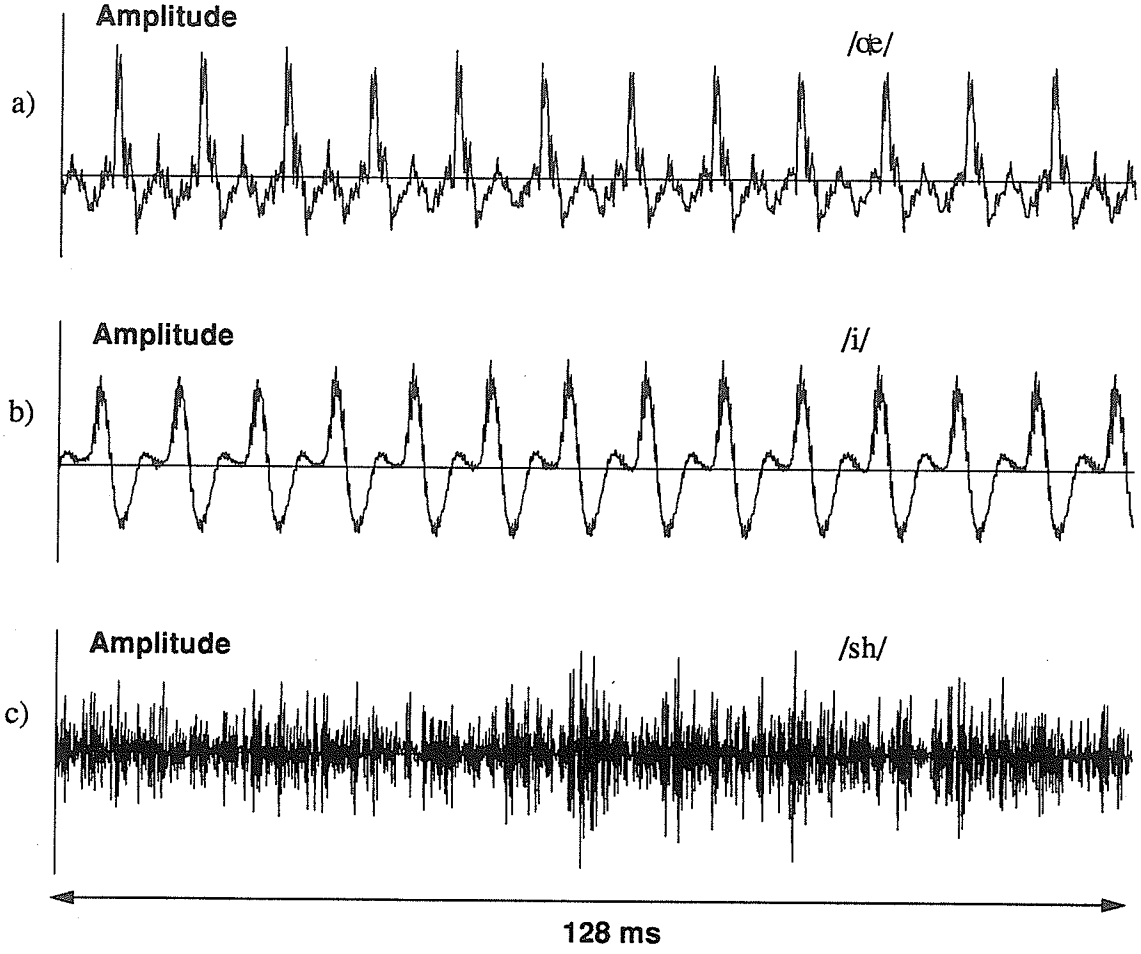


FIGURE 2. Waveforms for 128 ms segments from the words "hat", "beet", and "show".

Diphthongs account for the smooth transition from one vowel sound to another. For example, /aU/ as in "how" is generated by the transition from /a/ (hat) to /U/ (food).

Semivowels consist of the phonemes /w/, /l/, /r/, and /y/. These phonemes are difficult to classify because they are highly modified by the accompanying phoneme(s). However, they do have a vowel-like sound and thus the classification.

Consonants can be broken down into nasals, stops, fricatives, whispers and affricates. Nasals are voiced sounds produced by radiating from the nose instead of the mouth. Stops are shock waves generated with the lips, tongue or near the velum. They may be either voiced (/b/) or unvoiced (/p/). Whereas stops consist of short-time bursts of white noise, fricatives are excited by time-invariant white noise; for example /sh/ (show) as in Figure 2c. Voiced fricatives are excited by white noise generated at a constriction as well as a periodic glottal signal (for example, /v/ and /z/). Whispers are excited by white noise produced at the glottis. The nature of the whisper is usually determined by the following phoneme which determines the shape of the vocal tract. Affricates are created by following a stop with a fricative. An example is /j/ which is produced by following /d/ with /zh/.

By trying out all the phonemes and reflecting on the type of excitation, status of the velum, and shape of the vocal tract, one can gain very useful insight into the production of speech.

2.3 Analysis of the Vocal System

2.3.1 Voiced vs Unvoiced Speech

In the case of voiced speech the glottis stimulates the vocal tract with a periodic signal whose period is called the pitch period. Tension in the vocal chords causes the glottal opening to close more tightly. The result is that when air is forced through the opening, the glottis vibrates with a higher pitch.

A voiced sound has several peaks, or "formants", in its frequency response. Formant frequencies are determined by the resonances of the vocal tract. Figure 3 illustrates the Fast Fourier Transform (FFT) of the phoneme /œ/. The formant frequencies are 220 Hz, 690 Hz, 1.8 kHz, 2.6 kHz, and 3.6 kHz. Notice that signal levels above 4 kHz are more than 40 dB lower than the peak level (level of the first formant). This is the case generally with voiced sounds; unvoiced sounds, however, have appreciable signal content above 4 kHz.

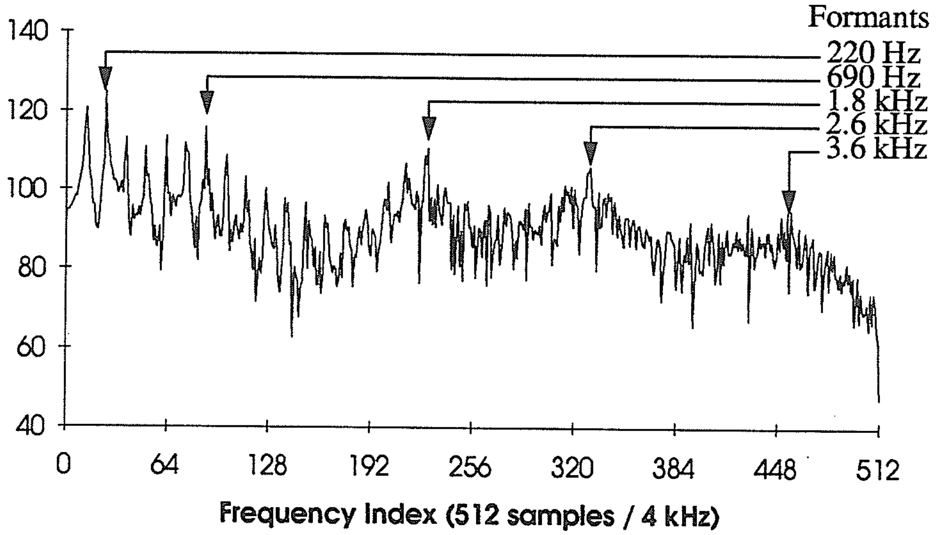


FIGURE 3. FFT of the phoneme /œ/ (hat). The peaks in the FFT correspond to the resonances of the vocal tract and the frequencies are called "formants".

For unvoiced speech, the glottis forces air through a constriction at some point in the vocal tract which results in broadband noise excitation. In this case, air forced up the trachea passes through the glottis but the glottis does not vibrate.

2.3.2 Vocal Tract

The vocal tract is shown in Figure 1. For a typical human male the vocal tract is 17 cm long. Most sounds are made by closing the velum preventing the nasal cavity from participating in the dynamics of the system. In this case, air enters at the glottis and exits at the lips. The vocal tract can be represented by an area function which gives the variation in area along the tract. An example is shown in Figure 4. In the case of nasals, sound cannot radiate from the mouth because it is blocked (i.e.: by the lips for the nasal /m/). However, the velum is open and sound does travel up the nasal cavity and is radiated at the nose. So the vocal tract works in two modes: nasal or non-nasal.

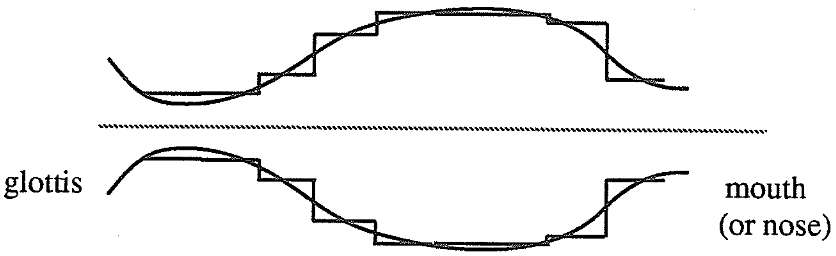


FIGURE 4. A typical vocal tract area function (smooth curve) and a concatenated-tube approximation.

The sound dynamics of the vocal tract can be modelled by two equations:

$$-\frac{\partial p}{\partial x} = \rho \frac{\partial (u/A)}{\partial t} \qquad (1)$$

$$-\frac{\partial u}{\partial x} = \frac{1}{\rho c^2} \frac{\partial (pA)}{\partial t} + \frac{\partial A}{\partial t} \qquad (2)$$

where *p(x,t)* is the variation in sound pressure with position *x* and time *t*, *u(x,t)* is the variation in volume velocity, ρ is the density of air in the tube, *c* is the speed of sound, and *A(x,t)* is the area function of the tube. Equation 1 expresses that a change in pressure across a distance will result in acceleration of the air molecules. This is illustrated in Figure 5.

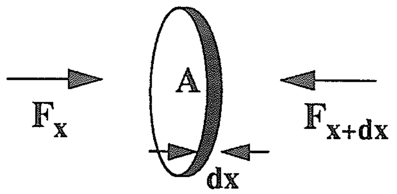


FIGURE 5. Forces due to pressure act on a disk of air inside the vocal tract.

The force acting on the disk of air is

$$dF = F_x - F_{x+dx} = Ap_x - Ap_{x+dx} = -Adp$$

The mass of the disk of air is $dm = \rho A dx$ and the acceleration can be written in terms of the volume velocity as $\boldsymbol{a} = \partial(u/A)/\partial t$. The bold quantities are vectors where a positive value indicates the direction towards the right in Figure 5. From Newton's law, $dF = dm\boldsymbol{a}$ we get

$$-Adp = (\rho A dx)(\partial(u/A)/\partial t)$$

which simplifies to Equation 1.

Equation 2 expresses the conservation of matter. The left hand side expresses the change in volume velocity with space. The second term on the right hand side is the change in area with respect to time. If the area is increasing with time, the volume velocity will decrease with space since some of that volume velocity is required to fill the opening space. The first term on the right hand side of Equation 2 is due to the compressibility of air and this term gives rise to the wave-like nature of the system.

Losses in the tube consist of viscous friction between the air and the walls of the tract, heat conduction through the walls of the tract, and vibration of the tract walls. These losses will result in frequency warping and attenuation of the frequency response of the tract.

Except for the most simple configurations, Equations 1 and 2 do not have closed form solutions. However, if we consider a general area function to be simplified to a series of uniform lossless tubes as shown in Figure 4, then for each tube the area function is constant for space and time and Equations 1 and 2 simplify to the following equation:

$$\frac{\partial^2}{\partial x^2}u(x,t) = \left(\frac{1}{c^2}\right)\frac{\partial^2}{\partial t^2}u(x,t) \qquad (3)$$

This is the wave equation and is easily solved given boundary conditions at each end of a tube. The boundary condition for the tube closest to the glottis is the glottal volume velocity excitation. The boundary condition for the tube at the other end of the vocal tract is the open air lip termination (discussed below). Adjacent tube ends must have equal volume velocity and pressure.

For voiced sounds, the glottal volume velocity is a periodic excitation signal. For unvoiced sounds, white noise excites the vocal tract, and the glottis can be considered to be the source of this noise to simplify the model.

An electrical analogy can be made if pressure is considered to be voltage and volume velocity is considered to be current. We can define an impedance to be pressure divided by volume velocity. In this case, the vocal tract can be modelled by a series of concatenated lines of characteristic impedance. The termination at the end of the line represents the lip termination and the glottal excitation is a current source. Rabiner and Schafer [1] have shown that a good approximation to the lip radiation impedance is the impedance characteristic of a plane baffle which is given by

$$Z_L(\omega) = \frac{P_{lip}(\omega)}{U_{lip}(\omega)} \approx \frac{j\omega L_r}{1 + j\omega L_r / R_r} \tag{4}$$

where $R_r = 128/9\pi^2$ and $L_r = 8a/3\pi c$. a is a free parameter used to best fit the model (see [1] for a more detailed explanation).

This model of the vocal tract is a linear model and can be approximated by a frequency response given by

$$\frac{U_{lip}(z)}{U_{glottis}(z)} = \frac{G}{1 - \sum_{k=1}^{P} \alpha_k z^{-k}} \tag{5}$$

where P is the number of poles. This all-pole frequency response sufficiently models the vocal tract when it has no energy sinks. However, for nasal sounds, zeros are introduced because energy can be trapped in the mouth cavity. Figure 6 shows the electrical analogy of the system for a nasal sound. The lips are closed resulting in zero volume velocity (current) which is analogous to an open circuit. Sound radiates from the nostrils which can be represented by an impedance similar to the lip impedance. In this case, the mouth cavity can trap energy leading to zeros in the frequency response of the vocal tract. This can be a problem when modelling the vocal tract since standard linear prediction techniques do not account for zeros in the system transfer function.

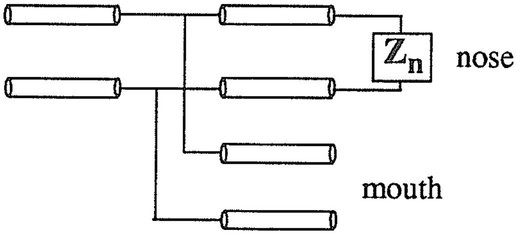


FIGURE 6. Electrical analogy of the vocal tract for a nasal sound. Z_n is the free-air interface impedance at the nose. Note that the line impedances vary with space.

As energy travels along the series of concatenated ideal tubes, it will be reflected at each boundary. Each reflection will delay the signal by 2τ, where τ is the time required for the sound to travel across one tube. Thus, the impulse response of the system will be of the form

$$h(t) = \sum_{k=0}^{\infty} \alpha_k \delta(t - 2k\tau - P\tau) \qquad (6)$$

where the output impulses occur every two time delays because of backward and then forward reflections. P is the number of tubes in the model (equivalent to the number of poles in the model), and the delay $P\tau$ in the output is the time required for the sound to travel from the glottis to the mouth or nose opening.

The Fourier transform of this impulse response is periodic in frequency with period $1/2\tau$. Thus, to prevent aliasing, the input to the model should be band-limited by $1/4\tau$; the sampling frequency should be $f_s = 1/2\tau$. The tube delay, τ, depends on the length of the vocal tract, l; the number of uniform tubes used to model the tract, P; and the speed of sound, c: $\tau = l/cP$. These relations lead to the following equation:

$$P = \frac{l}{c\tau} = \frac{1}{2\tau}\frac{2l}{c} = \frac{f_s}{1\text{kHz}} \qquad (7)$$

Values $l = 17$ cm and $c = 340$ m/s were used to obtain Equation 7. For a sampling frequency of 8kHz, the number of coefficients in the model should be at least 8. Fewer coefficients will cause aliasing in the vocal tract model.

Chapter 3: Frequency-Time Representation of Sounds

3.1 What Does Frequency-Time Mean?

As the vocal tract changes shape and the glottal excitation changes form, the frequency response of the vocal tract and the frequency content of the glottal excitation change. Fourier analysis proper does not explicitly account for the time-varying frequency content of a signal. This is not inherently problematic; however, when perceiving sounds, the human **does** break the signal down into time-varying frequency components via the cochlea.

In the time representation, a one-dimensional series of data expresses how the signal varies with time. In the frequency representation, we take the Fourier transform of the signal to get an expression of the signal as it varies in frequency. The frequency-time representation explicitly shows how the frequency components vary in frequency and time. Examples of the frequency-time representation are the short-time Fourier transform (STFT) and the wavelet transform.

All transforms discussed in this chapter are linear transforms. A useful paradigm for linear transforms is the Hilbert Space, H. A signal can be represented by a point in H. In the time domain, each dimension of H is an impulse at a particular time. In the frequency domain, each dimension is an impulse at a particular frequency. In the frequency-time domain, each dimension is a basis function which is in some way localized in both frequency and time. This allows the frequency-time representation to show how the signal varies in frequency and time.

3.2 Suitability of Frequency-Time Analysis to the Human Speech Channel

Physiological and neurological attributes of the cochlea and early auditory processing system of the brain indicate that sounds are broken down into frequency components changing in time by the cochlea and are processed in this way by the brain [3]. In essence, the cochlea performs a frequency-time transform of the auditory signal. So by considering a frequency-time representation of incoming sounds, we can understand how the cochlea and brain will transform them. For example, certain frequency channels in the cochlea may have a slower response time than other channels. Therefore the information rate in the slower channel is lower than the information rate in the faster channel. Such knowledge can be useful when choosing a coding scheme.

Sound vocalization can be modelled by either a periodic source or a white noise source being transformed by a time-varying linear filter which represents the vocal tract. A time-varying linear filter is most explicitly represented in the frequency-time domain. Again, certain frequency components of the time-varying transform may behave differently than other components and the frequency-time representation of the transform allows us to make use of this information.

Because of its physiological and neurological relevance, the frequency-time representation is very important in the area of non-linear transforms. An appropriate domain must be chosen before a non-linear method (i.e.: vector quantization) can be applied. For example, the compression technique of silence or zero detection would result in no compression of a pure sinusoid if applied in the time domain. However, if the zero detection transform were applied in the frequency domain, where the sinusoid is represented as a single non-zero element, excellent compression would result. The success of the frequency-time domain as a starting point for non-linear methods can be measured by experiment only. But the relevance of this domain to the vocalization and audition processes indicates that it may be a beneficial paradigm for speech compression.

3.3 Sound Spectrograph

One of the first devices developed to provide a frequency-time representation of a signal was the sound spectrograph. This device uses a bank of filters to show time-varying spectral characteristics. The two-dimensional map produced by the device is called a spectrogram. The vertical dimension corresponds to frequency and the horizontal dimension to time. The energy of the signal at a particular frequency and time is represented by the darkness of the map at that point. Fricatives contain broadband energy, whereas voiced phonemes contain bands in the spectrogram. Those bands correspond to the formants in the phoneme. It is possible to learn to read a spectrogram and determine the utterance from the spectrogram information alone. Note that the spectrogram is an electro-mechanical tool used for studying the time-varying frequency characteristics of sounds. The machine does not readily provide digital data. Also, the elements of the map do not correspond to the dimensions of a Hilbert space because they are not orthogonal. Since the spectrograph does not provide numerical values for the time frequency coefficients, we will not study how the spectrogram corresponds to a Hilbert Space. In later sections we will study how other transforms such as the Short-Time Fourier Transform (STFT) correspond to Hilbert Spaces.

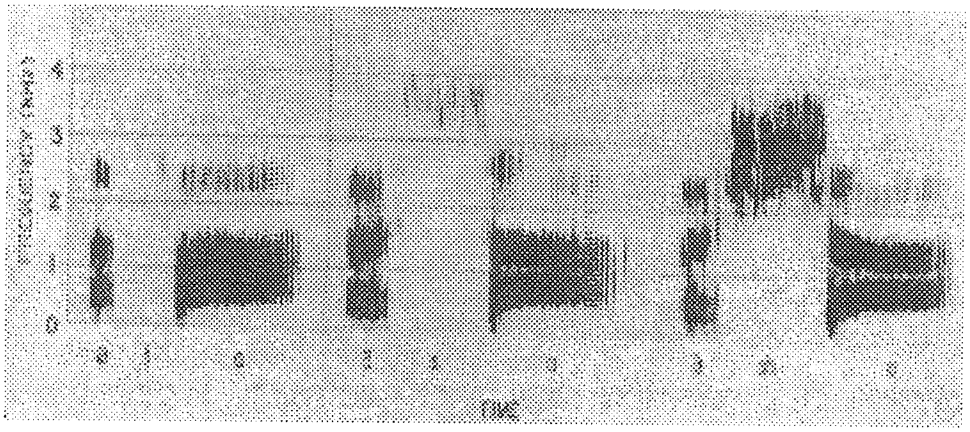


FIGURE 7. Sound spectrograph spectrogram of the utterances /UH-F-A/, /UH-S-A/, and /UH-SH-A/ from Rabiner and Schafer [1].

Figure 7 shows the sound spectrogram for three utterances. Notice that the fricatives, such as /s/, have energy spread across the spectrum, whereas voiced sounds, such as /UH/ have energy in frequency bands.

3.4 Short-Time Fourier Transform (STFT)

The method of short-time Fourier analysis provides a frequency-time representation of the signal by repetitively windowing the signal, taking the Fourier transform of the windowed signal, and shifting the window. The width of the window provides the localization in time for the Fourier analysis. The block diagram shown in Figure 8 indicates how the STFT is obtained. The resemblance of the transform to the Fourier transform with the addition of a window is obvious.

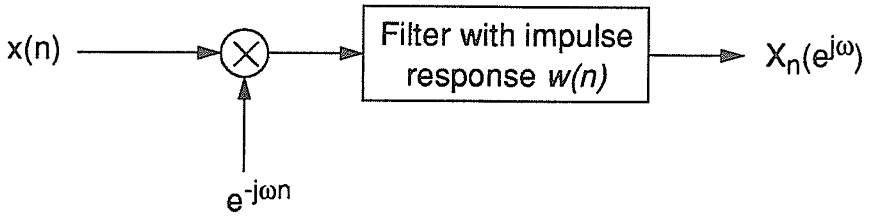


FIGURE 8. Block diagram of the STFT operation. Each signal sample is multiplied by a frequency-domain basis function as with the Fourier series. Then, a filter with a window-type impulse response is used to sum a portion of the weighted sample sequence. In this way, the frequency content of the signal in the locality of each sample is calculated.

The system equation is

$$H_n(e^{j\omega}) = \sum_{k=-\infty}^{\infty} w(n-k)x(k)e^{-j\omega k} \qquad (8)$$

The window *w(n)* selects the portion of the signal to be transformed.

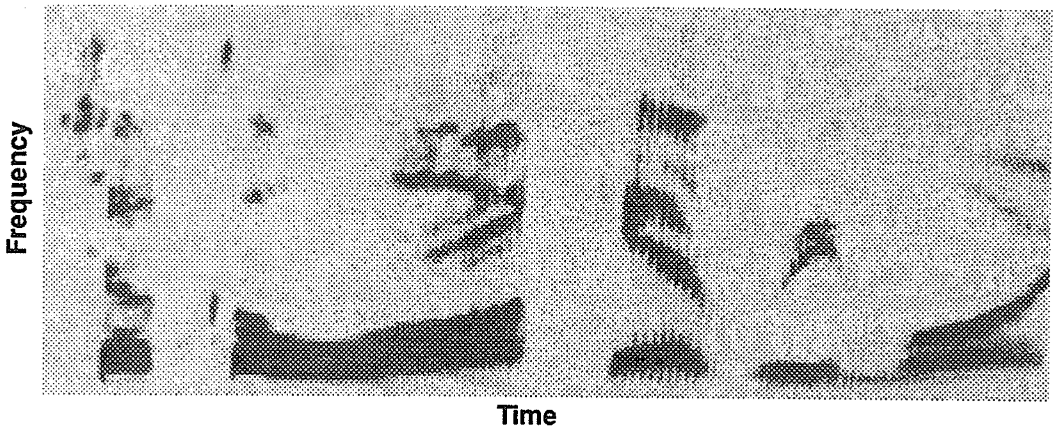


FIGURE 9. Computer-generated STFT sound spectrogram from Oppenheim [6]. It closely resembles the spectrograph output.

Figure 9 shows an example of the STFT in spectrogram format. Equation 8 can be written as a filter bank summation which indicates the resemblance between the STFT and the sound spectrograph (where the filter banks are implemented in hardware).

3.4.1 STFT Does NOT Have an Orthonormal Basis

When studying the frequency-time representation of a signal, it is important to know to what extent the different frequency-time states are orthogonal. For example, two adjacent non-zero coefficients may lead the observer to believe that the signal has more content than it does since the basis might be non-orthogonal. In addition, the entropy of the representation is low since information is duplicated by non-orthogonal states. Consequently, coding of the frequency-time representation must include careful consideration of the non-orthogonality.

The inner product between two bases indicates to what degree they duplicate information. We can examine the orthogonality of STFT states which differ in time but not in frequency by calculating the following inner product:

$$\begin{aligned}
\langle X_n | X_l \rangle &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{m=-\infty}^{\infty} w(n-m)x(m)e^{j\omega m}\right)\left(\sum_{k=-\infty}^{\infty} w(l-k)x(k)e^{-j\omega k}\right)d\omega \\
&= \frac{1}{2\pi}\int_{-\pi}^{\pi}\left(\sum_{m=k=-\infty}^{\infty} w(n-m)w(l-k)x(m)x(k)e^{-j\omega(k-m)}\right)d\omega \\
&= \frac{1}{2\pi}\left(\sum_{m=k=-\infty}^{\infty} w(n-m)w(l-k)x(m)x(k)\right)\int_{-\pi}^{\pi} e^{-j\omega(k-m)}d\omega \\
&= \left(\sum_{m=k=-\infty}^{\infty} w(n-m)w(l-k)x(m)x(k)\right)\delta_{km} \\
&= \sum_{m=-\infty}^{\infty} [w(n-m)x(m)]\,[w(l-m)x(m)]
\end{aligned} \tag{9}$$

The inner product of two states with same frequency and with different time indices depends on the window function and the signal itself. Consider the signal $x(m) = 1.0$:

$$\langle X_n | X_l \rangle = \sum_{m=-\infty}^{\infty} w(n-m)w(l-m)$$

This quantity will be zero only if the window function is uncorrelated with itself when shifted by (n-l). In general, then, the STFT does not have orthogonal states.

The fact that the STFT has a non-orthogonal basis does not mean that the signal cannot be recovered from the transform. In fact the conditions for recovery of the signal are reasonably weak: the window function at time 0 must be non-zero.

3.4.2 Phase Space Localization of Basis States

The location of the short-time Fourier transform basis states in phase-space (frequency-time space), is shown in Figure 10.

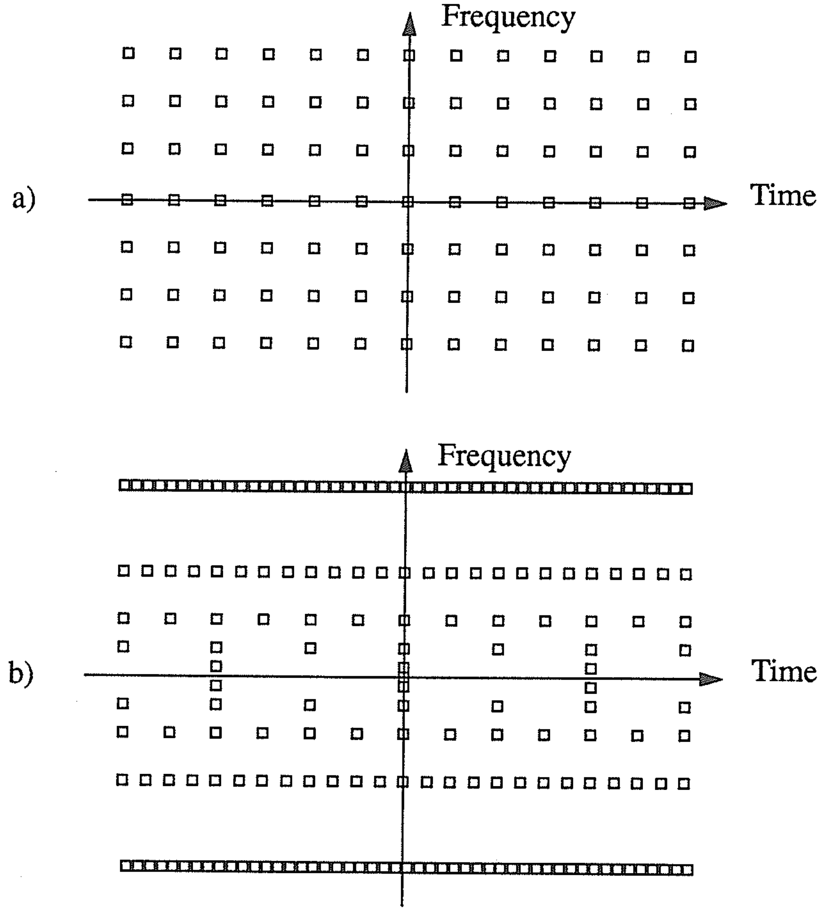


FIGURE 10. Location of STFT basis states (a) and wavelet transform basis states (b) in phase space (frequency-time space).

Notice that the lattice is regular, periodic in both time and frequency. This lattice illustrates a further drawback of the STFT: poor utilization of information. The zero frequency elements in the lattice are represented periodically in time even though they represent DC components of the signal. Obviously, the DC components of the signal do not vary with time, so these coefficients are wasted. High-frequency components have a short period and therefore can have an envelope which varies quickly with time. Consequently, a good representation of high frequencies should have states which are closer together in time than those at low frequencies. The STFT has a periodic lattice of states, not allowing for more resolved representation at high frequencies.

Gabor, who originally developed the STFT, used a Gaussian window. It can be shown that the Fourier transform of a Gaussian is a Gaussian. The Fourier transform of a Gaussian multiplied by the sinusoid of a particular frequency is a Gaussian shifted by that frequency. This may be the one saving grace of the STFT for it means that the basis states are well-localized in frequency and time.

3.5 Wavelet Transform

The wavelet transform, also called the "affine" transform, overcomes the STFT's inability to utilize information efficiently. In addition, a wavelet basis may be chosen such that it is an orthogonal basis. Consequently, it is possible to construct a wavelet basis which forms the basis for a Hilbert space in which almost any signal can be represented (the exceptions are non-Lebague-integrable functions which are indeed exceptional). The wavelet transform consists of projecting the signal on to the basis. This frequency-time representation of a signal is much simpler than the STFT since different coefficients correspond to orthogonal bases. Each coefficient indicates signal content that is independent of other coefficients.

3.5.1 Generation of the Basis from the Mother Wavelet

The transform name "wavelet" or "affine" refers to the way in which the lattice of bases in phase-space is assigned. "Affine" refers to the transform t' = at + b, where t, for example, is the time variable. Instead of modulating a Gaussian by sine waves at different frequencies, the affine transform warps and shifts the time variable of a basis state called the "mother wavelet". In this way, the entire basis is formed. As a result, the mother wavelet is "squeezed" or "stretched" and shifted in time to form the basis. The basis states are shifted by a smaller amount of time if squeezing occurs; they are shifted by a larger amount of time if stretching occurs. The assignment of basis states in phase space can be seen in Figure 10.

High-frequency wavelets are shifted less in time. This allows for a more efficient representation since high-frequency signals can have a more quickly changing envelope. This is what allows the wavelet transform to represent signals more efficiently than the STFT.

3.5.2 A Simple Example: the Haar Wavelet

A type of wavelet which clearly illustrates how wavelets work is the Haar wavelet. The mother wavelet from which all other wavelets are derived is called the Haar function:

$$h(t) = \begin{cases} 0, t < 0 \\ 1, 0 < t < 0.5 \\ -1, 0.5 < t < 1 \\ 0, 1 < t \end{cases} \tag{10}$$

The set of basis states are generated from the affine transform as follows:

$$h_{m,n}(t) = 2^{m/2}h(2^m t - n) = 2^{m/2}h(2^m\left[t - n2^{-m}\right]) \tag{11}$$

Increasing m leads to more squeezing; decreasing m leads to more stretching. The second form in Equation 11 indicates that highly squeezed wavelets are closer together in time. Figure 11 illustrates the mother wavelet in addition to other wavelets generated from the affine transform.

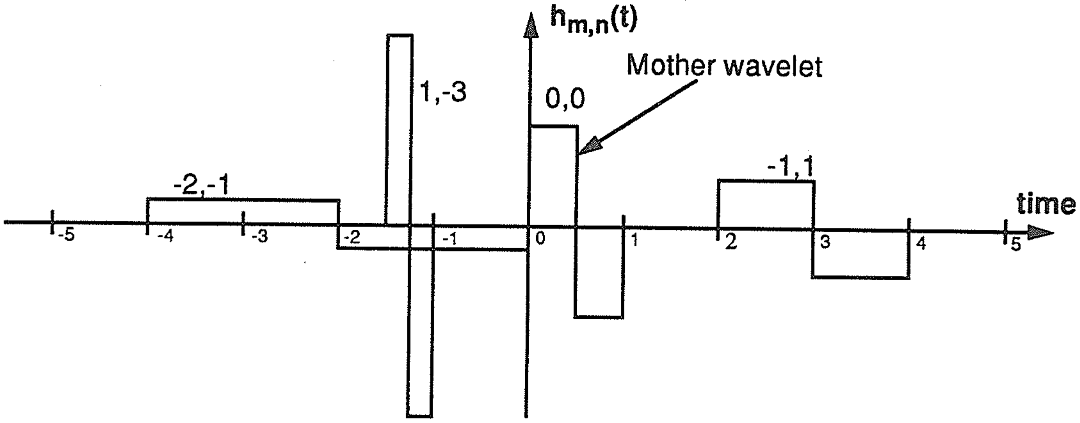


FIGURE 11. The wavelet basis is formed by shifting and then stretching or squeezing the mother wavelet. The scale for shifting is determined by the stretching or squeezing, so that high-frequency wavelets are shifted more finely. The values for m,n are shown beside each wavelet.

Consideration of how the wavelets are generated indicates that the basis is orthogonal. The affine transform squeezes/stretches a wavelet by a factor of two and shifts it by the width of the generated wavelet. The inner product of these two wavelets will consist of the inner product of an odd function with the even portion of another function which is zero. Wavelets that have not been stretched or squeezed with respect to each other will be shifted in time by an amount such that the wavelets do not overlap and, of course, the inner products will be zero. Therefore, this wavelet basis is orthogonal.

By mixing various wavelets together, it is possible to obtain a time-domain impulse function located at any particular instant. An infinite number of wavelets will be needed, but the sum will have vanishingly small coefficients. Although a rigorous proof is not presented here, one can become convinced that this is possible simply by trying out combinations. It follows that if any impulse function can be expressed as a weighted sum of wavelets, the wavelets form a basis for $L^2(\Re)$. Since the Euclidean norm of each wavelet is unity, the Haar wavelets presented form an orthonormal basis.

3.5.3 Daubechies Discrete Wavelet Transform (DWT)

Although the Haar wavelet transform has an orthonormal basis and has good time localization of the basis states, it does not have good frequency localization of the basis states. As seen in Figure 13a, the FFTs of the various basis states overlap significantly. Daubechies [5] has developed a variety of wavelet transforms. Already very popular are her discrete wavelet transforms (DWTs) [7], which are generated from a set of wavelet filter coefficients. The specific way in which the wavelet transform matrix, **D**, is constructed from the wavelet filter is presented in Appendix B. The final result is

$$DWT(\mathbf{s}) = \mathbf{D}\mathbf{s} \tag{12}$$

The sample vector, **s**, is linearly transformed by the transform matrix. Because the pyramidal algorithm can be used when applying the transform, the calculation is order *N*, where *N* is the number of samples. (Note: *N* must be an integer power of 2.)

It is easiest to understand the DWT by considering each row of the transform matrix to be a basis state vector. The sum of two such basis states are shown in Figure 12 for the case of Daubechies' 20-coefficient DWT. They obviously have different time and frequency indices. This wavelet basis is orthonormal, has good time localization, and has good frequency localization.

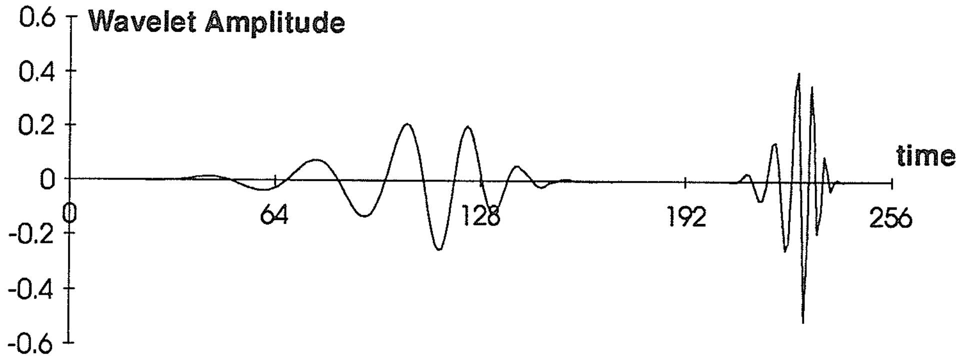


FIGURE 12. Time-domain basis states for the DWT generated from Daubechies' 20-coefficient wavelet filter.

The frequency localization of the Haar wavelet basis states and the so-called DAUB20 wavelet basis states are shown for comparison in Figure 13.

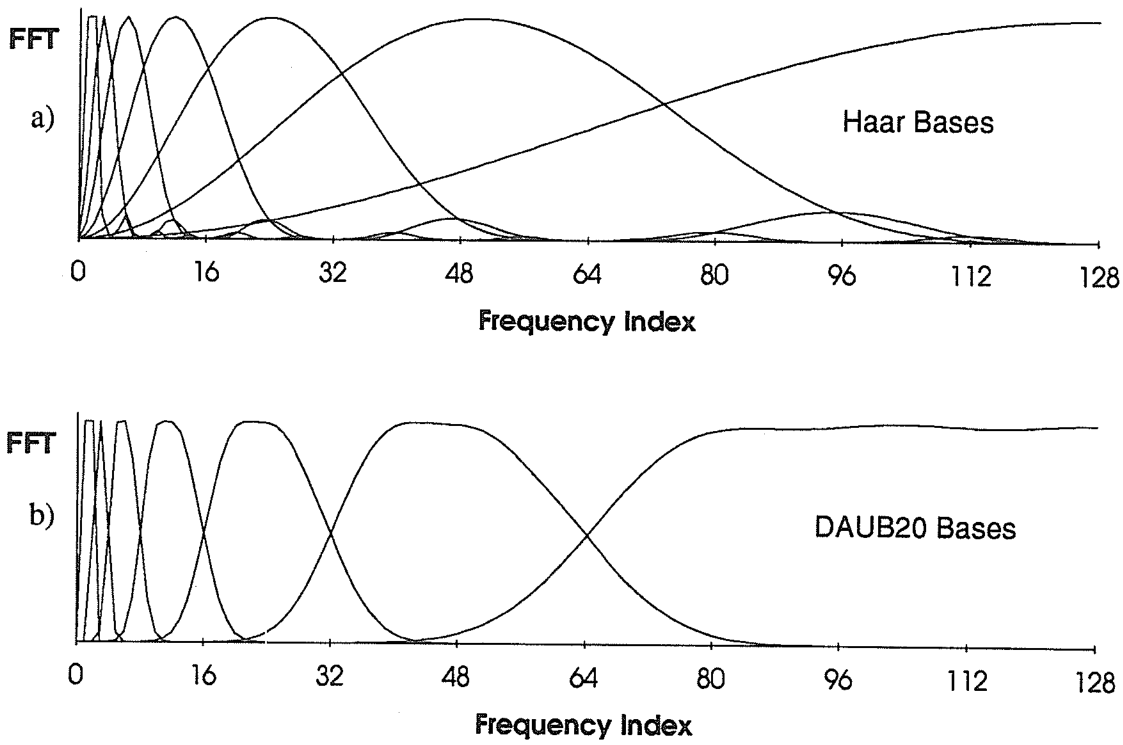


FIGURE 13. The FFT of the Haar basis states (a) indicates a greater amount of overlap in the frequency domain than for the DAUB20 basis states (b).

A plot similar to a spectrogram can be used to show the DWT of a signal (Figure 14). Instead of having uniform resolution, however, the DWT has varying resolution: fine at high frequencies; rough at low frequencies.

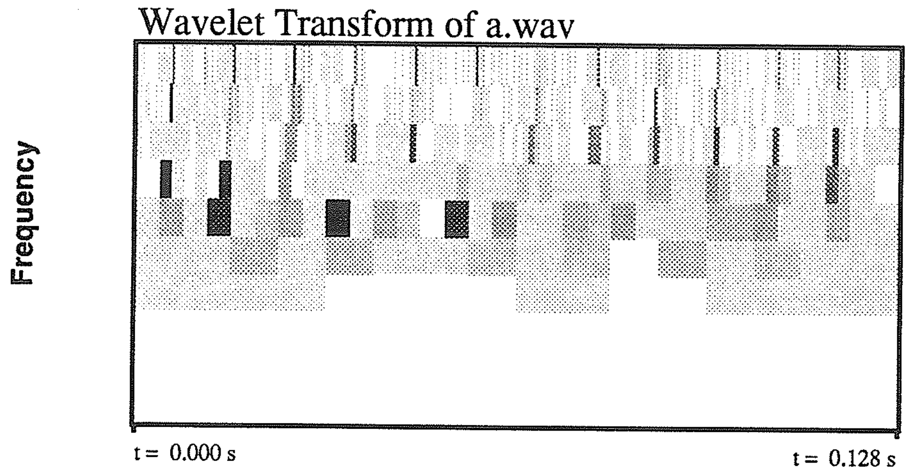


FIGURE 14. The DWT of /æ/ (hat). Notice the finer temporal resolution at higher frequencies.

According to the uncertainty principle, decreasing the localization width of one variable causes an increase in the localization width of its conjugate variable: $\Delta t \Delta \omega \geq 1/2$. The superior temporal localization of the Haar wavelet leads to inferior frequency localization. A Daubechies DWT which has bases which are more highly localized in time than the DAUB20 wavelets are the DAUB4 wavelets, shown in Figure 15.

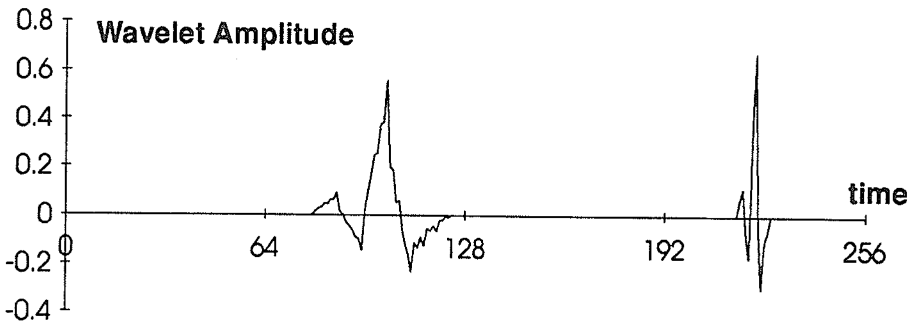


FIGURE 15. Time-domain basis states for the DWT generated from Daubechies' 4-coefficient wavelet filter.

Chapter 4: Representation of Speech

In Section 2.3.1 we saw that for voiced sounds, frequencies above 4 kHz are more than 40 db attenuated from the peak amplitude. However, for unvoiced sounds such as /s/, there is appreciable signal content above 4 kHz. However, practical implementations of telephone systems have proven that 4 kHz of bandwidth is sufficient for good quality speech transmission [8]. This means that a rate of 8000 samples / second (Nyquist rate) is suitable for sampling the speech signal for telephone quality sound.

The technique of sampling a signal, coding (quantizing) the pulses, and reconstructing the signal from the pulse codes is called PCM (Pulse Code Modulation). "Code" means that a code is sent down the line to reconstruct the pulse level at the sampling rate.

4.1 Optimal PCM

Shannon's information rate provides a good measure of compression. If 1 of N symbols is sent over the channel at a time, Shannon's information rate or entropy measure is:

$$H = -K\langle \log p_i \rangle = -K \sum_{i=1}^{N} p_i \log p_i \tag{13}$$

where p_i is the probability of symbol i occurring. If K is chosen as follows, the entropy measure can be used to compare the information utilization of different sets of symbols:

$$H = -\frac{1}{\log N}\langle \log p_i \rangle = -\frac{1}{\log N}\sum_{i=1}^{N} p_i \log p_i \qquad (14)$$

An entropy of 1 indicates good utilization of the symbol set. An entropy of 0 indicates extremely poor utilization. For the case of PCM, the set of symbols is the set of quantization levels. For 8-bit PCM, there are 256 symbols or quantization levels. Figure 16 shows the code utilization for 8-bit uniform PCM applied to a variety of both male and female human speech. Uniform quantization does not give uniform usage of the PCM codes. In fact, the entropy of the code book is quite low: 0.59. Paez and Glisson [9] have shown that a good approximation to the speech amplitude distribution is a gamma distribution. If the quantization levels are assigned according to this distribution, then code usage will be uniform, giving an entropy near 1.0. Such an assignment of quantization levels would result in a clustering of levels near 0 (fewer levels at high amplitude).

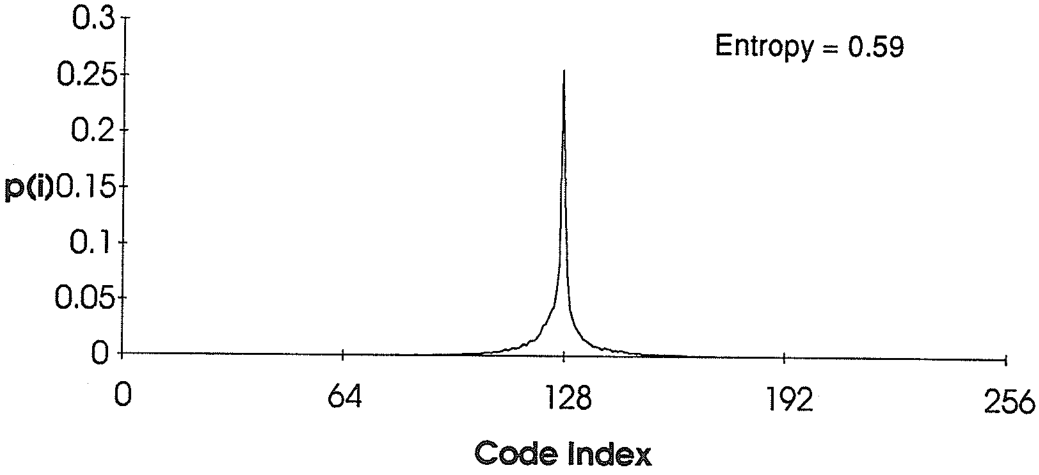


FIGURE 16. Code utilization for 8-bit uniform PCM. This curve shows the probability distribution for the uniformly assigned codes.

Another consideration when assigning quantization levels is the quantization error. Indeed the optimal assignment of levels so as to maintain a constant percentage quantization error is a logarithmic distribution [1]. The principles of high entropy codebook utilization and uniform quantization error are important in all areas of coding, including vector quantization. Later, Shannon's entropy measure will be used to measure coding performance.

Using the above ideas, a coding scheme called μ-law has been developed [10]. In this case the following equation is applied before uniform PCM.

$$y(k) = x_{max} \frac{\log\left(1 + \mu \frac{|x(k)|}{x_{max}}\right)}{\log(1+\mu)} \operatorname{sgn}[x(k)] \tag{15}$$

This coding scheme provides telephone quality speech at a sampling rate of 8000 bytes per second or 64000 bits per second [1].

4.2 Wavelet Representation

In Chapter 2 and Chapter 3 we saw that the dynamics of speech are of a frequency-time nature. Therefore, it makes sense to look for coding schemes in the frequency-time domain. Techniques such as the short-time Fourier transform have been used to compress speech [11]. However, the wavelet transform's ability to zoom in on short-time features has generated interest in compressing speech using DWTs [12].

Figure 15 illustrates the DWT of the vowel in "hat". The temporal periodicity in the DWT indicates the pitch period of the voiced sound. The time-domain signal remains near zero most of the time so that the DWT is sparse. A sparse representation of a signal is ripe for compression.

4.3 Wavelet Transform Compressed

An implemented zero-detection coder breaks the speech signal into 16 ms segments. For each segment, it calculates the DWT, quantizes the transform coefficients to a number of levels (either 8 or 16) and transmits the location and quantization level of only the non-zero components. This scheme generates only slightly distorted speech at an average rate of 12800 bits per second (1600 bytes per second).

Since the vocal tract changes fairly slowly with time, the frequency response of a vocal tract is approximately constant during the pitch period. From one pitch period to the next, the glottal excitation is usually similar (except when switching from one phoneme to another). A technique called Linear Predictive Coding (LPC) makes use of this (see the following section). However, during changes of the glottal excitation, short-time features are present which may not be represented by the simple coding methods used in LPC. However, these features can be efficiently coded using the DWT (Chapter 5).

4.4 Linear Predictive Coding (LPC)

For the linear predictive coding of speech, we assume that the glottal excitation and vocal tract have a simple model. The vocal tract can be modelled by an all-pole linear filter (Equation 5). If we assume that the glottal excitation is either a series of impulses (whose period is the pitch period) or stationary white noise, then the particular shape of the actual glottal excitation can be modelled by adding poles to the vocal tract filter. Atal and Hanauer [13] have found that the addition of four poles is sufficient. From Equation 7, the total number of poles needed to model the "vocal tract" for LPC is

$$P = \frac{f_s}{1kHz} + 4 \tag{16}$$

For 8000 sample /second input speech, 12 poles are needed.

This model does not include the zeros of the vocal tract transfer function (mainly caused by the mouth during nasal sounds) nor the non-linear effects of the vocal system. We also assume that the true glottal waveform can be sufficiently simplified by the 4-pole glottal filter. For each segment of speech, usually about 30 ms in duration, the filter coefficients are calculated; it is determined whether the sound is voiced or unvoiced; if the sound is voiced, the pitch period is calculated; and the gain of the system is found. Given these parameters, either random white noise for unvoiced sounds or an impulse train with period determined by the pitch period for voiced sounds is applied to the linear filter given by the filter coefficients. This gives a perceptual approximation to the original speech segment. Note that although the perceptual error may be low, the Euclidean error between the two waveform vectors will be very high. Figure 17 illustrates the synthesis block diagram.

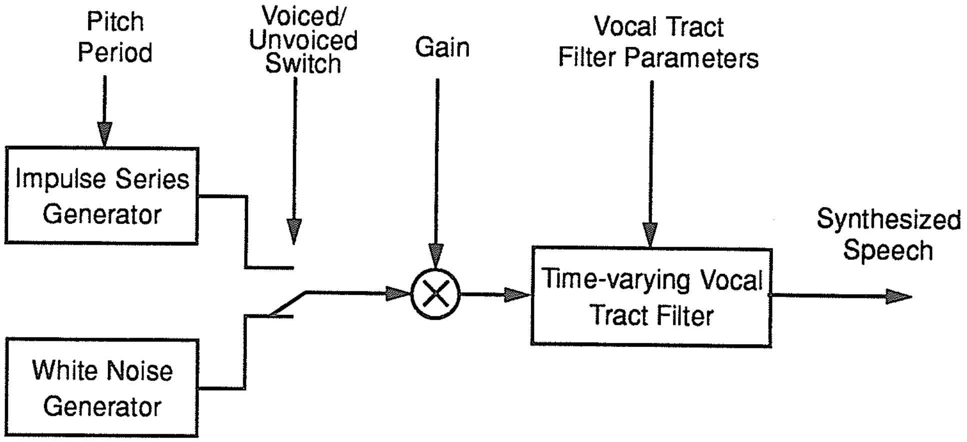


FIGURE 17. Block diagram of the LPC synthesizer.

If *u(m)* is the synthetic excitation, then the synthetic speech samples *s(m)* are generated (see Equation 5) by

$$s(m) = Gu(m) + \sum_{k=1}^{P} \alpha_k s(m-k) \tag{17}$$

It can be shown that if *u(m)* is a periodic impulse series or white noise, then the technique of linear predictive analysis gives the correct coefficients $\alpha_1, ..., \alpha_P$[1][7]. The linear predictive analysis technique used for this thesis is Burg's method (see Appendix C).

The LPC technique was implemented under DOS on a 20MHz 386 PC. Filter parameters are coded using 4 bits each, the voiced/unvoiced switch is coded using 1 bit, the pitch period is coded using 5 bits, and the gain is coded using 6 bits. The LPC parameters are calculated every 25 ms giving a total bit rate of 2400 bits / second.

Figure 18 shows the FFT of the phoneme /æ/ from Figure 3 and the FFT of the LPC synthesized phoneme. The spectra match reasonably well. Notice that the formant peaks are preserved.

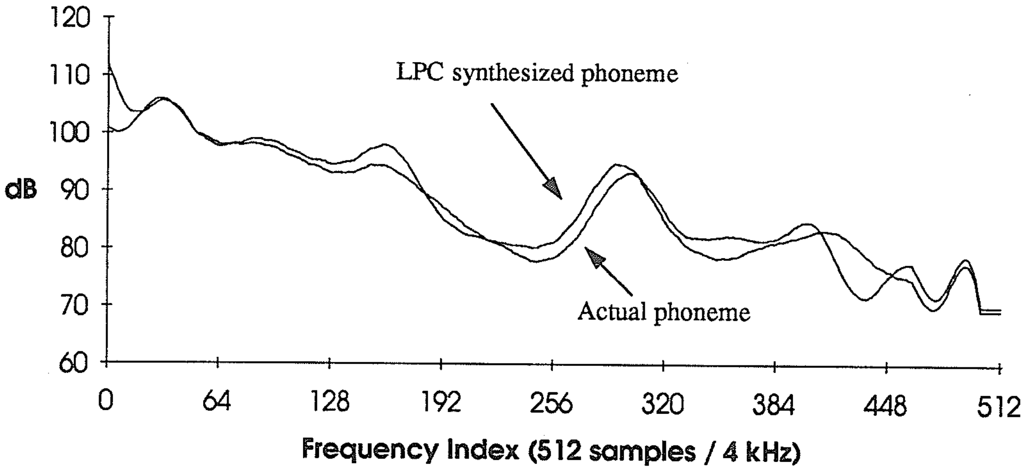


FIGURE 18. Comparison between the FFT of an actual phoneme, /æ/ (hat) and the FFT of its LPC synthesized counterpart. The formants are preserved.

So LPC efficiently determines the vocal tract transfer function. However, in many cases, LPC generates erroneous codes because the excitation signal does not follow its assumed form or the filter does not follow its assumed form (such as for nasals). Also, stop constants, such as /p/, have short-time dynamics which are smeared out by the frame analysis of LPC. Figure 19 shows the DWT of the original utterance /p/ and of the LPC synthesized sound.

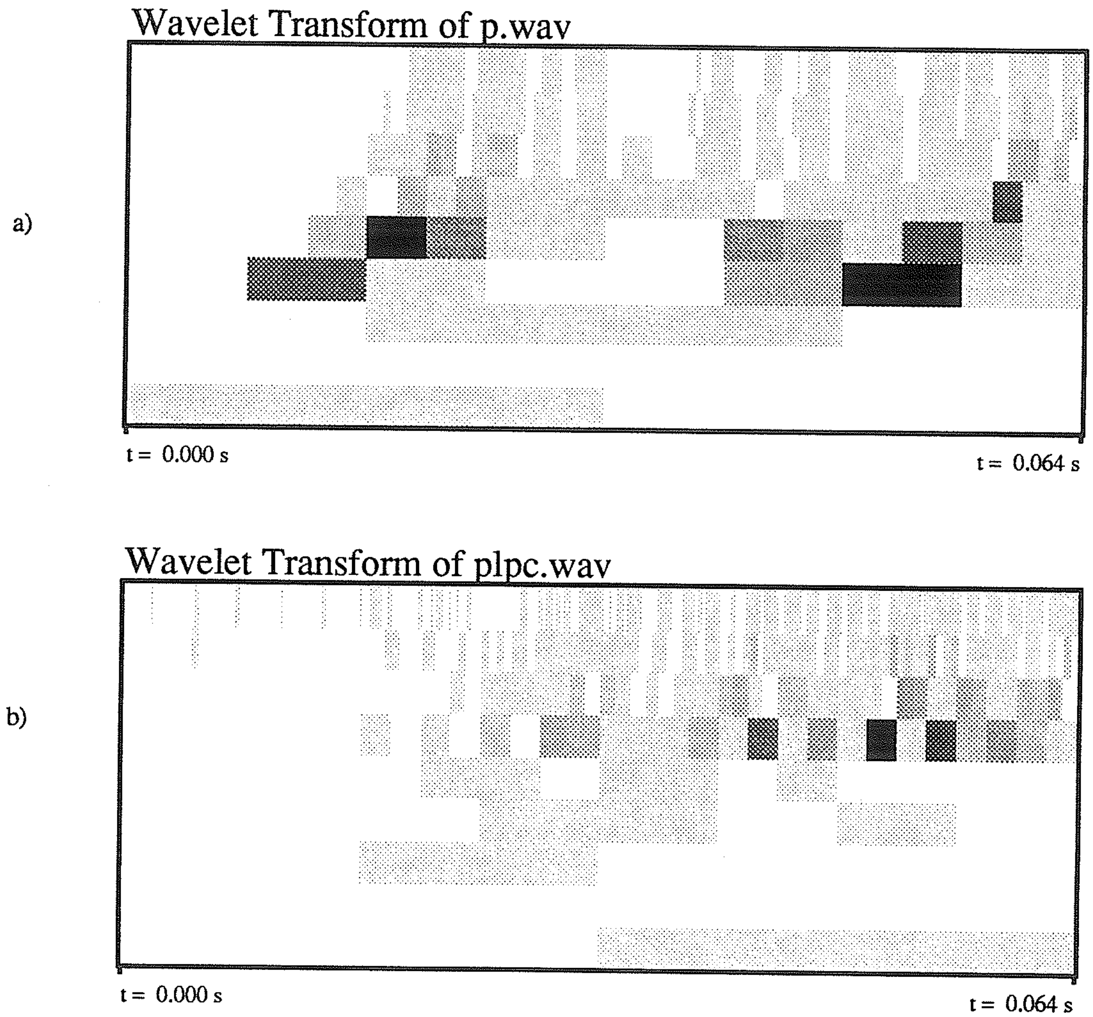


FIGURE 19. DWTs of the original utterance of /p/ (a) and the LPC synthesized utterance (b). The frame loses the short-time features of the stop consonant.

4.5 Code Excited Linear Prediction (CELP)

4.5.1 True Excitation: More Than White Noise or Impulses

An LPC vocoder synthesizes speech by assuming the excitation is a series of impulses or a white noise source. If the vocalization system were ideal, this would indeed be sufficient.

Figure 20 illustrates the excitation obtained by applying the inverse linear predictor filter to the voiced phoneme from which the predictor was obtained. The excitation is seen to be more complex than a series of impulses.

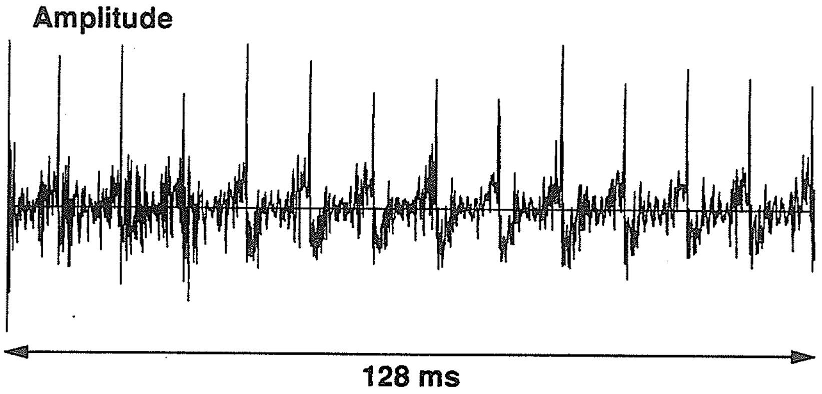


FIGURE 20. By applying the inverse vocal tract filter to the phoneme /œ/, we see that the true excitation is more complex than a series of impulses.

4.5.2 Using a Code to Approximate the True Excitation

If instead of using an impulse train or a white noise source we use a code selected from a codebook to excite the linear predictive filter, we obtain Code Excited Linear Prediction (CELP). Various methods have been used to determine the optimal code. Different application domains place different restrictions on the technique used. Some of the application issues are low-delay, low bit rate, high noise tolerance, and, of course, cost [14].

4.5.3 Important: Pre VS Post Filter Error

If we consider the block diagram shown in Figure 17, there are two obvious ways of obtaining an error measure for true speech and given synthetic excitation:

1. The inverse vocal tract filter can be applied to the real speech, generating a true excitation. The Euclidean (sum-of-squares) error between the **true excitation** and the **synthetic excitation** is then calculated.
2. The vocal tract filter can be applied to the synthetic excitation, generating synthetic speech. The Euclidean error between the **true speech** and the **synthetic speech** is then calculated. This method is called “analysis-by-synthesis”.

2 makes the most sense as a perceptual error, since it is the final output that the listener hears. However, 1 makes sense from the point of view of conservative computation.

4.5.4 Analysis by Synthesis CELP

Traditionally, CELP coding is based on analysis-by-synthesis techniques. The following describes a basic CELP vocoder [2] whose block diagrams are shown in Figures 21 and 22.

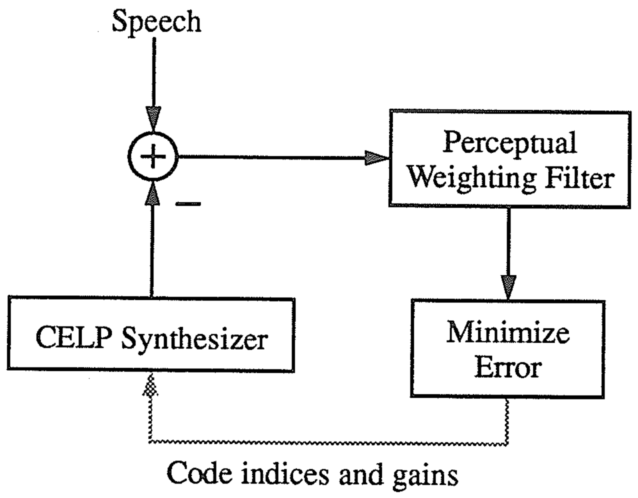


FIGURE 21. Block diagram of the CELP Analyzer.

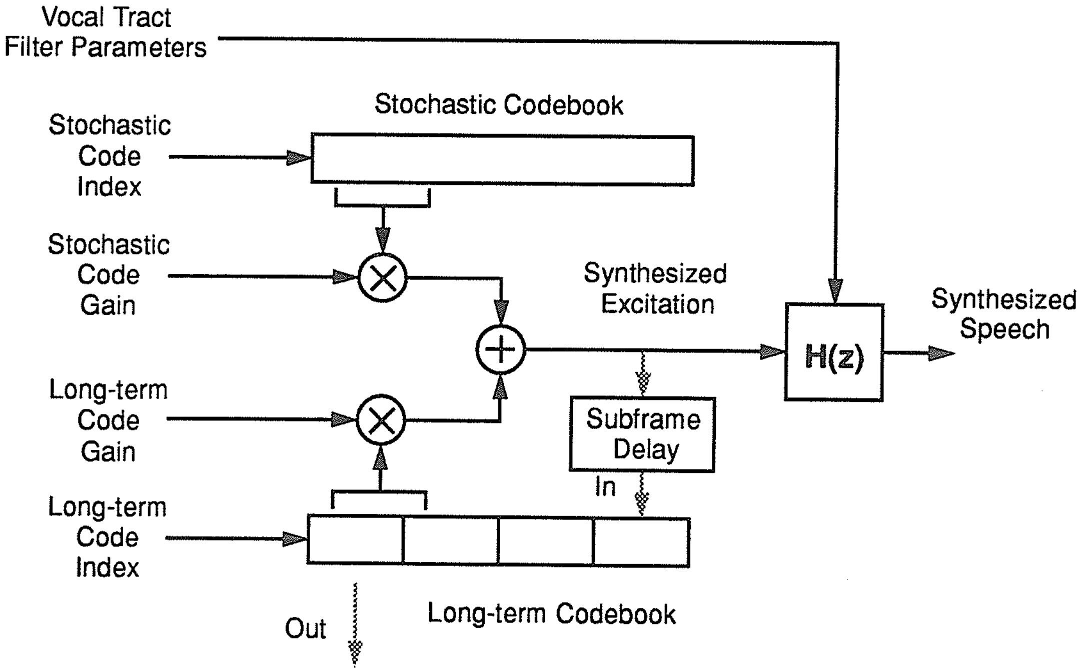


FIGURE 22. Block diagram of a CELP synthesizer.

For analysis, a code is selected, the linear predictive filter is applied, and the output is subtracted from the real speech to give an error. This error is then fed through a perceptual waiting filter and the code which minimizes this error is chosen to be the best code. The perceptual weighting filter enhances certain components of the signal so as to make the error minimization procedure relevant to the perceptual domain. Figure 21 illustrates the block diagram for the CELP analysis procedure. The CELP analyzer must synthesize the speech segment for each code to be tested.

Synthesis usually uses two (or more) codebooks in order to account for different properties of the true excitation. Figure 22 shows the CELP synthesizer.

Usually the code used to excite the prediction filter consists of two components. One is a long-term predictive code which is generated by shifting previous excitation codes into a buffer called the "long-term codebook". The long-term codebook may contain four or more subframes of previous excitation. The current code from the long-term codebook is selected by choosing a section of the codebook that minimizes the perceptually weighted error. Added to the long-term code is another code from a stochastic codebook. The long-term codebook contains long-term periodicity of the voice waveform, whereas the stochastic codebook allows for unpredictable excitation spikes as well as innovation when new sounds begin. It is a fixed codebook which contains random numbers which satisfy a particular distribution.

The subframe size is 64 samples and since the long-term codebook stores four subframes, the long-term codebook is 256 samples long. The stochastic codebook is 1024 samples long. Linear prediction occurs at multiple subframe intervals. For example, four subframes comprise a frame; the linear predictive coefficients are calculated for each frame whereas the codebook indices and gains are calculated for each sub-frame. The linear predictive coefficients may be interpolated for each subframe.

For each subframe, an optimal adaptive codebook index and gain and a stochastic codebook index and gain must be determined. Usually this is done using analysis by synthesis, which means the excitation is applied to the filter and an error is generated which is then perceptually weighted by another filter.

Although the analysis-by-synthesis CELP vocoder requires a large amount of computation [2], synthesized speech quality is much better than for LPC.

An additional advantage of the CELP vocoder is that it does not presume a particular type of excitation. Consequently, it is more capable of handling non-vocalized sounds than the LPC vocoder. This is useful for faithfully reproducing background noise.

Chapter 5: Code and Wavelet Excited Linear Prediction (CWELP)

5.1 Introduction

The advantage of the LPC vocoder is that it uses linear prediction which does not require much computation. Its disadvantage is that the excitation is not complicated enough to account for nonideal excitation. The advantage of the analysis-by-synthesis CELP vocoder is that it can adaptively create more complicated excitation to improve synthesized speech quality. Its disadvantage is the high computational requirement for the analysis-by-synthesis techniques. It seems that a compromise might be a good solution to the speech coding problem. One idea is to try to match an excitation code to the signal generated by applying the inverse predictive filter to the speech segment. This has the advantage that the analysis by synthesis stage is skipped (i.e. the linear predictive filter and perceptual weighting filter need not be applied) and this significantly reduces the computational power required. At the same time, a more complicated excitation signal can be generated leading to sound quality which is more robust than that from the LPC vocoder. The price to be paid for the model is higher computational complexity than the linear predictive vocoder and lower speech quality than the CELP vocoder (see Section 4.5.3).

5.2 CWELP System

Figure 23 illustrates the synthesis block diagram for the vocoder described above. Both short-term and long-term signals are generated and summed together to give the synthesized excitation signal. As with CELP synthesis, this excitation signal is delayed and added to the long-term codebook. The linear predictive filter is applied to produce the synthesized speech. However, as the analysis block diagram in Figure 24 shows, the error is calculated earlier than in the CELP vocoder. Here, we try to obtain the optimal excitation instead of the optimal speech signal output.

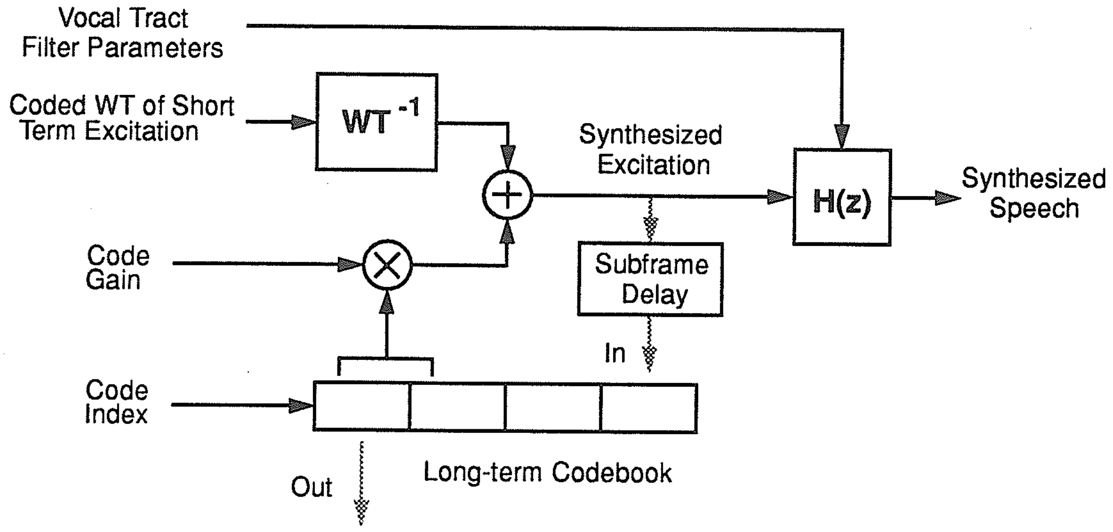


FIGURE 23. Block diagram of the CWELP synthesizer.

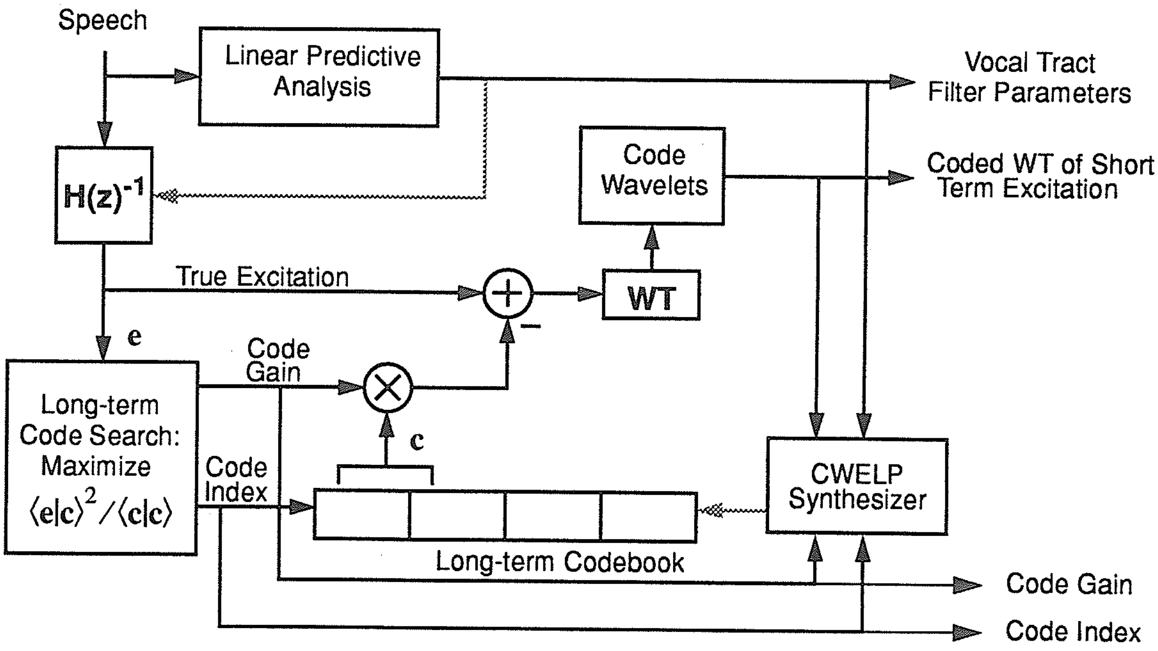

Speech
Linear Predictive Analysis
Vocal Tract Filter Parameters
H(z)-1
Code Wavelets
Coded WT of Short Term Excitation
True Excitation
WT
e
Code Gain
Long-term Code Search: Maximize ⟨e|c⟩²/⟨c|c⟩
c
Code Index
CWELP Synthesizer
Long-term Codebook
Code Gain
Code Index

FIGURE 24. Block diagram of the CWELP analyzer.

Since the short-term excitation is used to represent spontaneous spikes and innovate for when new sounds begin, most of the time the short-term excitation is zero. **The wavelet transform is an excellent way of representing high-frequency short-time spikes** [5][7]. Other short-term excitation compression schemes were not investigated; the wavelet transform domain was chosen for its ability to zoom in on oscillatory features (see Section 3.5).

The method used to determine the short-term excitation is to calculate the linear predictive coefficients, apply the inverse linear prediction filter to the speech segment to generate a true excitation signal, select an optimal amplified code from the long-term codebook and subtract it from the true excitation signal. This short-term excitation vector is transformed by the wavelet transform and is then compressed in the wavelet domain. The long-term codebook index and gain and the compressed wavelet transform of the short-term residual is transmitted through the channel to synthesize the speech. Since the excitation is generated from a series of codebook codes and wavelets, the compression technique is called the Code and Wavelet Excited Linear Predictive vocoder or CWELP vocoder.

The CWELP vocoder uses 8,000 sample/second input. Each frame consists of 256 samples and the linear prediction coefficients are calculated once per frame. The order of the linear predictor which minimized model error and bit rate was determined experimentally to be 10. A long-term codebook code index and gain and the coded wavelet transform of the short-term excitation are transmitted once per sub-frame (there are four subframes per frame).

5.2.1 Qualitative Difference Between Short-term and Long-term Codes

During voiced speech segments on the order of 20 ms duration, a significant portion of the vocal tract excitation energy is in the form of a periodic signal (at the voiced pitch). The long-term codebook is a means of storing recent excitation with the hope that future excitation will be similar. If the true excitation were as simple as a periodic series of impulses, the long-term codebook would be sufficient to perfectly reproduce the excitation.

Innovative excitation features (not accounted for by recent excitation) must be accounted for by a means other than the long-term codebook. Indeed, if these features were not coded, the long-term codebook would have no way of incorporating recent excitation approximations. These innovative excitation features are considered to be the "short-term" excitation.

5.3 Long-term Codebook

The size of the long-term codebook is 256 samples. Each code consists of a 64-sample segment of the codebook, so that each code overlaps by 63 samples with its adjacent codes. 8 bits are sufficient therefore to give the position of the code in the long-term codebook. Searching for the optimal code can be considered to be finding the optimal match between a codebook vector and a 64-sample segment of the true excitation. If **e** is the excitation vector and **c**(i) (where i ranges from 0 to 255) are the codebook vectors, the optimal codebook vector is found by maximizing the squared inner product of the excitation vector **e** with the normalized code vectors $\mathbf{c}(i)/\|\mathbf{c}(i)\|$. Using the squared inner product saves computation required by the square root operation and allows for negative gain (when the unsquared inner product is negative). α, the code vector gain, is then found by dividing the unsquared inner product by the squared norm of the code vector as shown below. Figure 25 illustrates the scenario used to derive Equation 18 and Equation 19.

$$i_{opt} = \left\{ i \,\middle|\, \left\langle \mathbf{e} \middle| \frac{\mathbf{c}(i)}{\|\mathbf{c}(i)\|} \right\rangle^2 \geq \left\langle \mathbf{e} \middle| \frac{\mathbf{c}(j)}{\|\mathbf{c}(j)\|} \right\rangle^2 \forall j \right\}$$

$$= \left\{ i \,\middle|\, \frac{\langle \mathbf{e}|\mathbf{c}(i)\rangle^2}{\langle \mathbf{c}(i)|\mathbf{c}(i)\rangle} \geq \frac{\langle \mathbf{e}|\mathbf{c}(j)\rangle^2}{\langle \mathbf{c}(j)|\mathbf{c}(j)\rangle} \forall j \right\} \qquad (18)$$

$$\alpha = \frac{\langle \mathbf{e}|\mathbf{c}(i_{opt})\rangle}{\langle \mathbf{c}(i_{opt})|\mathbf{c}(i_{opt})\rangle} \qquad (19)$$

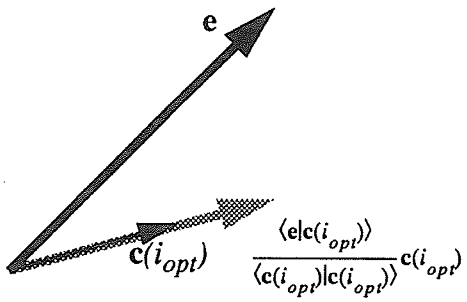


FIGURE 25. Projecting the excitation onto the code.

Once the complete excitation is found (including the short-term excitation; see below) the subframe of synthesized excitation is shifted into the long-term codebook buffer. The oldest sub-frame in the buffer is lost. Consequently, the long-term codebook consists of the last four sub-frames of synthesized excitation. It is from this that future long-term codes can be selected. Figure 26 shows the excitation for the phoneme /æ/, along with the long-term code excitation. Note that at first, the long-term codebook does not contain the information to account for the periodicity of the excitation. However, the short-term excitation synthesizer provides the innovation, such that the long-term code excitation eventually does mimic the periodicity of the true excitation.

The long-term codebook is restricted to a size of 256 samples for two reasons. First, extending the length of the codebook increases the codebook search time. Since the codebook search requires the greatest portion of CWELP analysis (Section 5.7), it is desirable to keep the codebook as small as possible. Second, 256 samples correspond to a period of 256 / (8000 Hz) = 32 ms. It is reasonable to assume that most ordinary speech will have pitch periods below this limit. Since only one pitch period need be stored in order to select the optimal short-term excitation segment for the next subframe, a codebook size of 256 is sufficient.

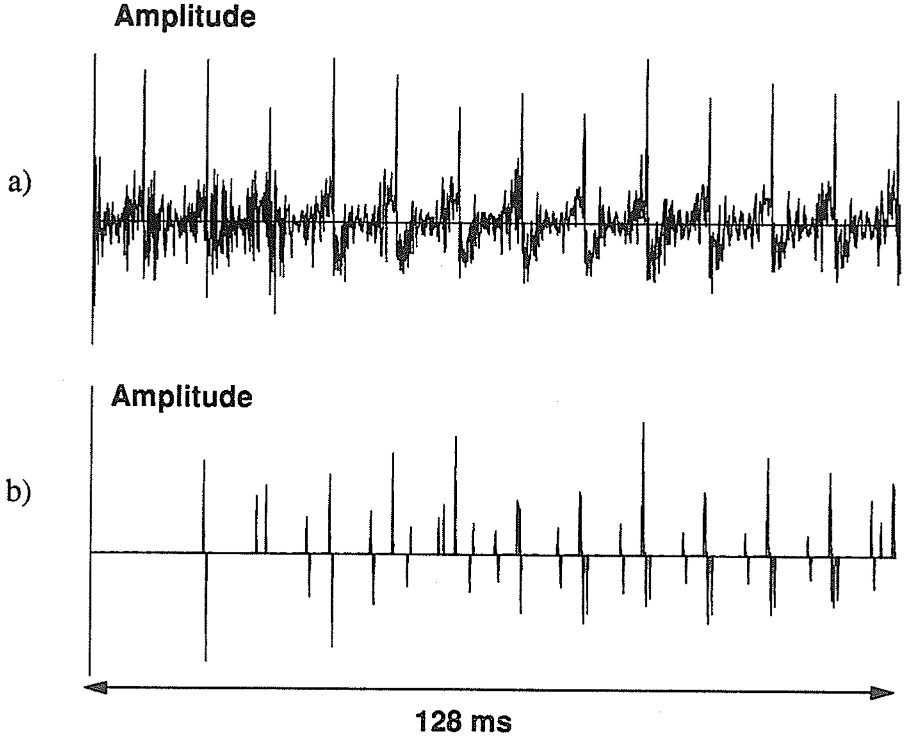


FIGURE 26. True excitation (a) and long-term codebook excitation (b). With time, the codebook stores useful information about the excitation.

5.3.1 Coding the Code Index and Gain

Figure 27 shows the probability distribution of the code index determined by applying the CWELP analyzer to a signal consisting of concatenated utterances of letters from the alphabet. The utterances were extracted from the ISOLET database (Appendix D) and were spoken by both males and females.

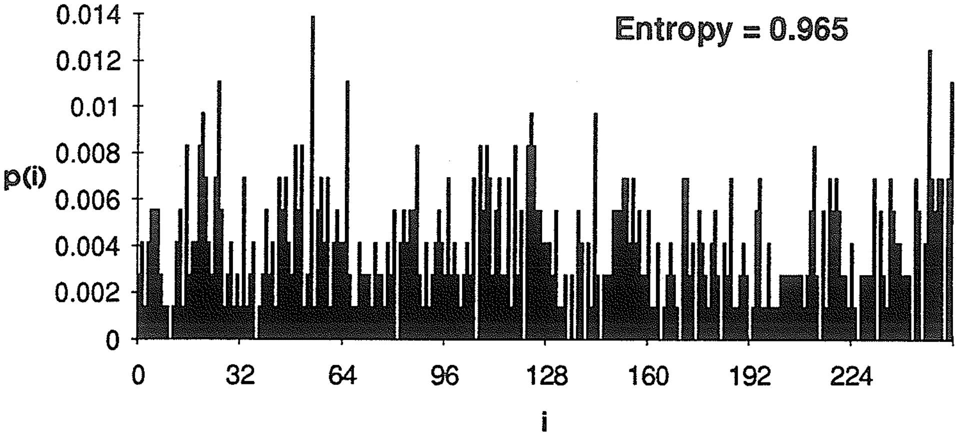


FIGURE 27. Probability distribution of the code index, i.

Evidently, the set of code indices are well utilized: the distribution has an entropy of 0.965.

Figure 28 shows the probability distribution of the absolute code gain. A distribution model that fits the data is

$$p(x) = 6.67x^2e^{-x^2/0.45} + 0.01, 0 \le x \le 10 \tag{20}$$

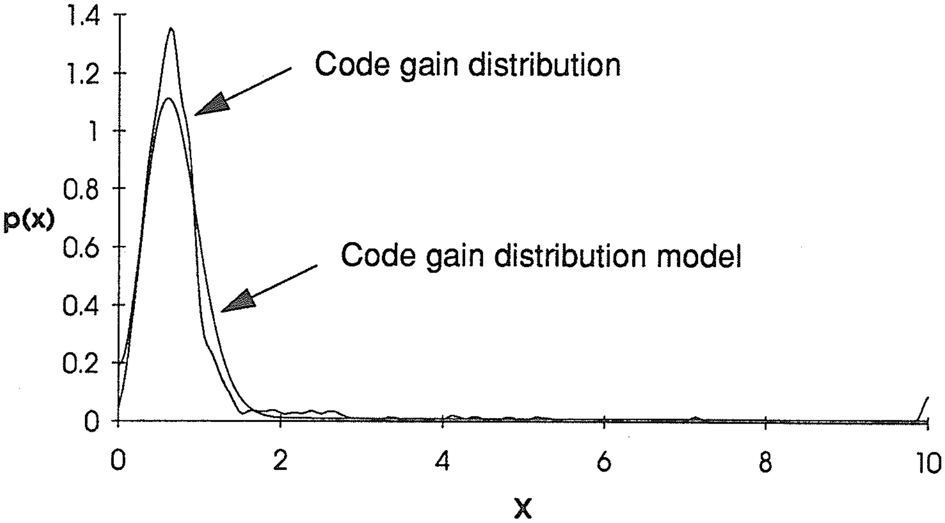
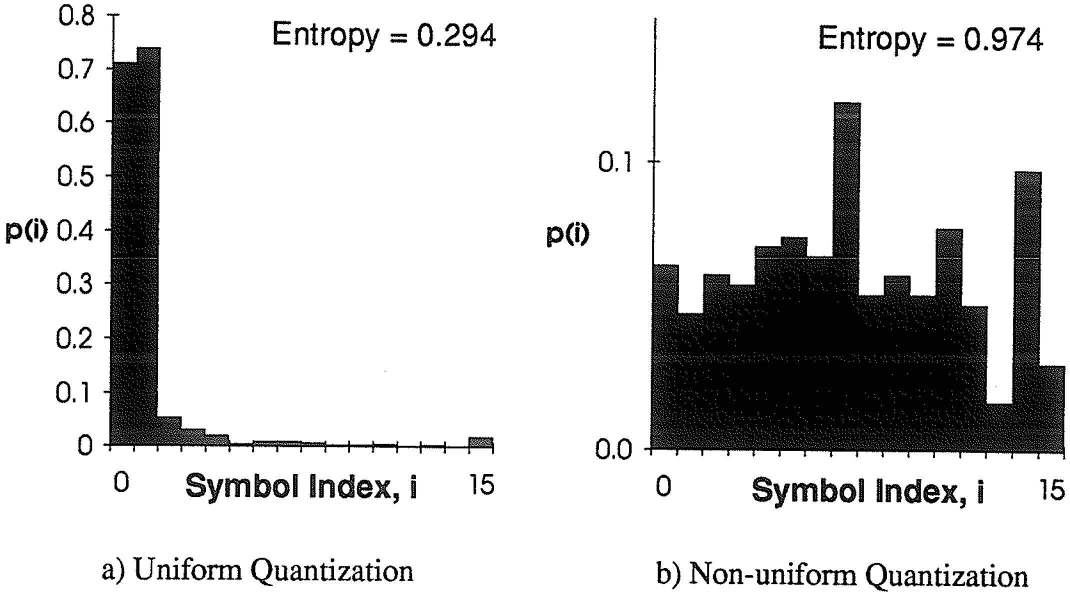


FIGURE 28. Probability distributions of the absolute code gain (a) and the code gain distribution model (b), where x is the absolute code gain.

The second term in the sum is necessary to account for gains greater than 2.0 which typically occur following a segment of silence (when the near-zero codebook codes must suddenly account for a sizeable signal). The distribution of the actual gain (as opposed to the absolute gain) is mostly even, indicating that the distribution of the absolute gain is a good approximation to the distributions for the negative and positive gains.

It was experimentally determined that 5 bits of non-uniform quantization are adequate to code the code gain without degradation in speech quality. One bit is used to represent the sign of the gain, while the other 4 bits represent the quantization level index of the absolute gain. The non-uniform quantization levels are assigned by dividing the code gain distribution model (Figure 28) into 16 sections of equal area. Consequently, the likelihoods of occurrence of the quantization levels should be almost equal, giving an entropy near unity. Placing a quantization level at the centroid of each section ensures that the mean square quantization error within each section is minimized.

Figure 29 shows the symbol utilization (2^4 symbols) for uniform quantization and non-uniform quantization of the code gain for test data. Whereas the uniform quantization scheme has a distribution entropy of 0.294, the non-uniform quantization scheme has a distribution entropy of 0.974.



a) Uniform Quantization

b) Non-uniform Quantization

FIGURE 29. Symbol utilization for the uniform quantization scheme (a) and the non-uniform scheme (b).

Since the index is coded in 8 bits, and the gain is coded in 5 bits, the long-term codebook excitation requires 8+5 = 13 bits per subframe, or 4x13 = 52 bits per frame.

5.4 Short-term Excitation

Figure 30 shows the wavelet transform of the short-term excitation (Figure 31) which is generated by subtracting the long-term code from the true excitation (Figure 26). The short-term excitation consists of short, high-frequency bursts. The wavelet transform is well known for its ability to efficiently represent such signals [4]. Compared to the time-domain signal, the wavelet transform gives a much simpler representation (i.e. fewer of the components are large).

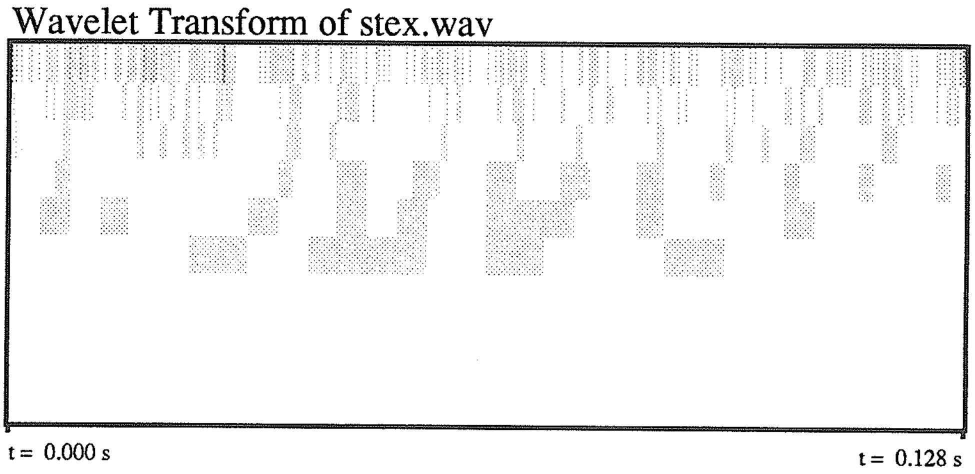


FIGURE 30. Wavelet transform of the short-term excitation. The excitation (Figure 31) consists of many short-time spikes and it is well-suited to compression in the wavelet domain.

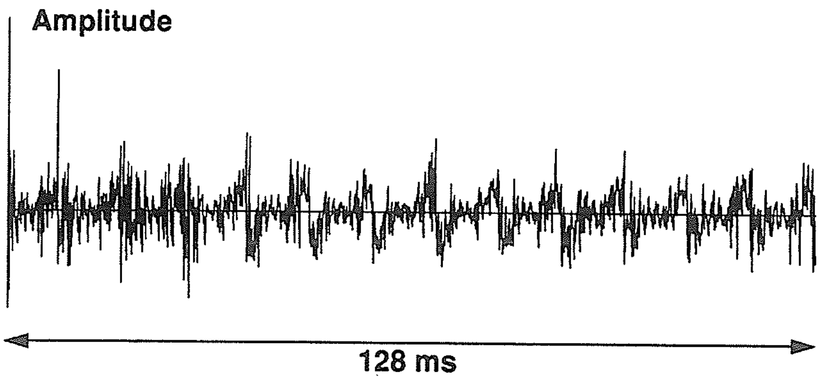
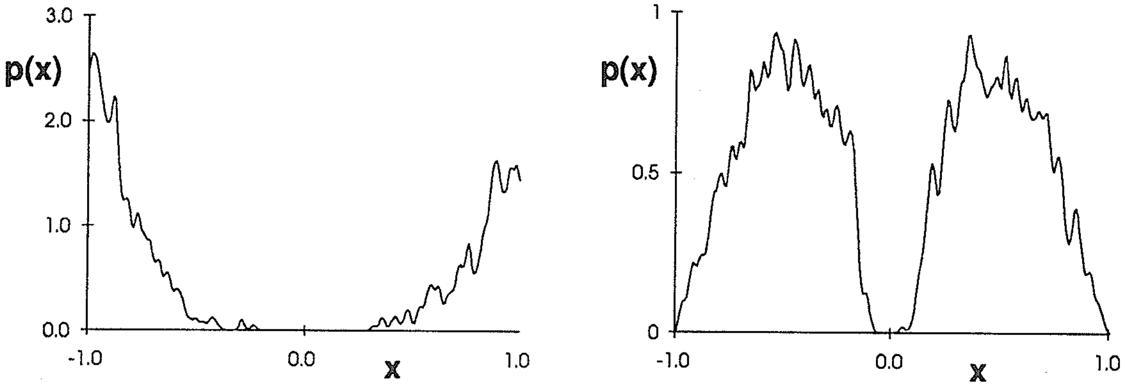


FIGURE 31. Short-term excitation (true excitation - long-term codebook excitation).

5.4.1 Coding the Short-term Excitation

Two coefficients per sub-frame seem to be adequate to account for the short-term innovation of the excitation. Figures 32a and 32b show the probability distribution of the second-largest wavelet coefficient and the cumulative distribution of the five next largest coefficients, averaged over a database (Section 5.3.1) of excitation signals minus long-term codes. Notice that the coefficient range is [-1,1]: the largest coefficient is used to normalize the magnitudes. The second-largest coefficient is almost always of nearly the same magnitude as the largest coefficient, whereas the next five coefficients drop significantly in magnitude. Of even less importance are the remaining coefficients.



a) Second-largest

b) Five next largest

FIGURE 32. Amplitude probability distributions of the second-largest (a) and the five next largest (b) wavelet coefficients, where x is the coefficient value.

As with the absolute code gain, the distribution in Figure 32a is assumed to be even, and four quantization levels (2 bits) are assigned non-uniformly to maximize the entropy of the symbol distribution.

Only the 32 highest frequency wavelet coefficients (uppermost band in Figure 31) are considered during coding; the first 32 wavelet coefficients do not seem to be important for maintaining reproduced speech quality. Also, since the long-term codebook accounts for periodicity, accurate timing is not necessary for the short-term excitation. Consequently, the 32 high-frequency coefficient indices are quantized to 16 indices (4 bits).

Four bits code the magnitude of the largest wavelet coefficient in a manner similar to the coding of the absolute code gain. One bit codes the sign of the largest coefficient. Four bits code the positions of each of the two largest wavelet coefficients. Two bits code the relative magnitude of the second largest coefficient. Therefore, the wavelet transform is coded in 4+1+4+4+2 = 15 bits per subframe, or 4x15 = 60 bits per frame.

Figure 33 shows the short-term excitation and the coded short-term excitation. Although most information is lost, the coded short-term excitation does account for spikes and some of the second order effects (usually caused by damped oscillations in the vocal tract [1]).

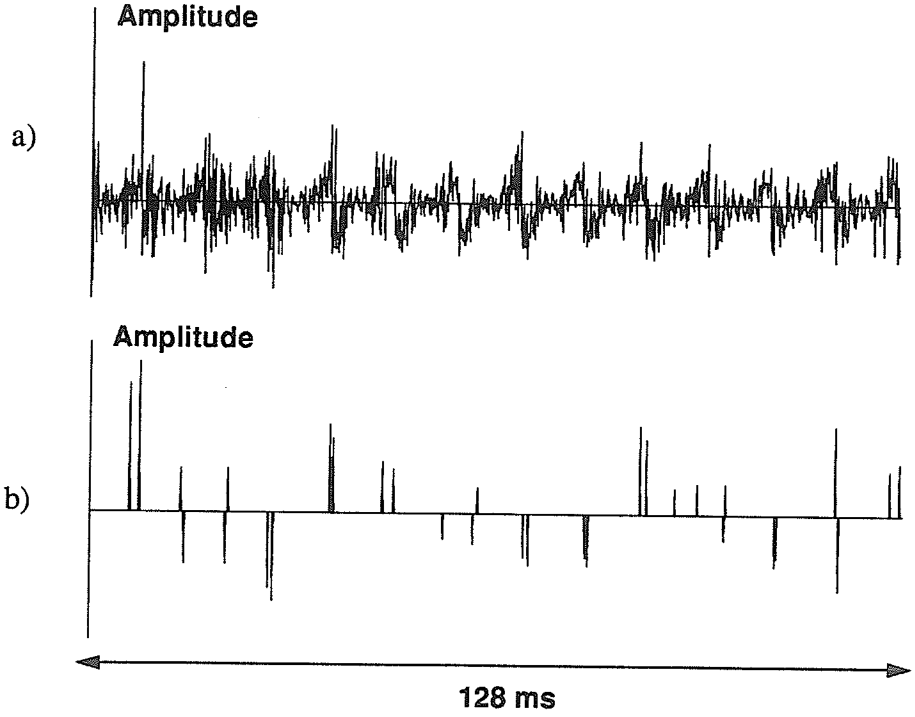


FIGURE 33. Short-term excitation (a) and the coded short-term excitation (b).

5.5 Discussion of Synthesized Excitation

Figure 34 shows the true excitation and the synthesized excitation for the phoneme /œ/. Innovation by the short-term excitation quickly allows the long-term codebook to account for periodicity. The periodicity is well-modelled; however, the synthetic excitation does not have the same envelope as the true excitation. This gives rise to a "roughening" effect in the synthesized speech. The sound level does not vary smoothly from one pitch period to the next.

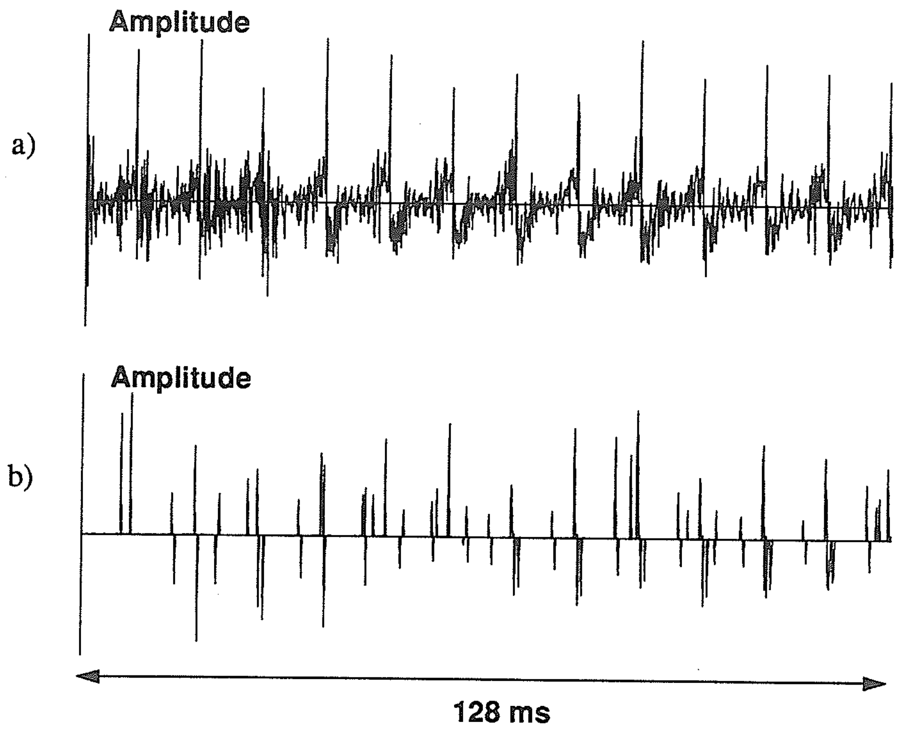


FIGURE 34. Comparison of the true excitation (a) and the synthesized excitation (b).

5.6 Bit Rate

Since 10th order linear predictive analysis is used, 10 vocal tract coefficients must be transmitted for each frame. The partial correlation coefficients [1] are suitable for transmission since they are bounded by the interval (-1, 1) in order to maintain stability and since they show very low inter-parameter correlation. These PARCOR coefficients are non-uniformly scalar quantized; the set of vocal tract parameters are coded in 34 bits. Table 1 summarizes the allocation of bits per frame.

TABLE 1. Summary of bits transmitted per CWELP frame

	Number of Bits	Percentage
Linear Prediction	34	23%
Long-term Codebook	4 x 13 = 52	36%
Short-term Coded Wavelet Transform	4 x 15 = 60	41%
Total	**146**	**100%**

Since 146 bits are transmitted per 256 sample frame, and since the input rate is 8000 samples / second, the CWELP vocoder has a bit rate of 146 x 8000 / 256 = **4563 bits/second.**

5.7 Computational Complexity

For the purpose of comparing vocoder computational complexity, the analyzer is of prime importance. The synthesizer usually requires a small fraction of the computation required by the analyzer (often the difference is greater than an order of magnitude). Table 2 shows the allocation of theoretical computation for the CWELP analyzer.

TABLE 2. Theoretical Computational complexity of the CWELP analyzer

	MIPS[a]	Percentage
10th order linear prediction	0.4	25%
Inverse vocal tract filter	0.08	5%
Long-term codebook search	1.04	63%
Wavelet transforms (2)	0.12	7%
Total	**1.64**	**100%**

[a] An "instruction" is considered to be a multiply or a multiply-and-accumulate, for the purpose of calculating MIPS (Million Instructions Per Second)

As with CELP, the majority of the analyzer's computation time is spent on codebook searching. An exhaustive search was used. The computation required by the long-term codebook search could be reduced by using other methods.

Chapter 6: Performance

6.1 Comparison of Bit Rates

Table 3 gives a comparison between the bit rates for LPC (Section 4.4), CWELP, and CELP (Section 4.5). Both CWELP and CELP have double the bit rate of LPC due to the excitation coding and transmission.

TABLE 3. Comparison of bit rates

	Bit Rate (bits/sec)
LPC	2400
CWELP	4563
CELP	4800

6.2 Comparison of Computational Complexity

One motivation behind the CWELP vocoder was to model more complicated excitations without introducing much more computation than the LPC vocoder requires (Section 5.1). At the same time, it was hoped that the CWELP vocoder's computational requirements would be much less than the CELP vocoder's. Table 4 gives a comparison between LPC, CWELP, and CELP.

TABLE 4. Comparison of computational complexity

	Theoretical MIPS	Measured MIPS
LPC	0.58	1.7
CWELP	1.64	3.6
CELP	12.6	25[a]

[a] This is the MIP rating of the DSP chip that Campbell et. al. [2] implemented their CELP vocoder on.

Both the LPC and CWELP vocoders were implemented on an IBM PC-compatible 20MHz 386; the algorithms did not run real-time, but the measured computation rates were adjusted to give the true rates:

$$\text{Measured-MIPS} = \frac{(\text{Coding-Time})\ (\text{MIPS-Rating-of-PC})}{(\text{Duration-of-Utterance})} \qquad (21)$$

As expected, the CWELP vocoder requires somewhat more computation than LPC but much less than CELP.

6.3 Recognition Tests Using the ISOLET Database

Five female and five male speakers were selected from the ISOLET Database (see Appendix D). One pronunciation of each letter of the alphabet was compiled for each speaker to give a total of (5+5)x26 = 260 test utterances. Although single-letter listening is not a usual test for speech compression schemes, this data was readily available and proved successful at eliciting recognition mistakes. Other types of tests used for vocoder comparisons are Dynastat's diagnostic rhyme test (DRT) and diagnostic acceptability measure (DAM) [15].

A total of 1332 utterances were read to two female and four male listeners. It was found that 8000 sample / second 16-bit PCM recordings were accurately identified 90% of the time. That 10% of the uncompressed recordings were misidentified indicates that the data set was noisy to begin with. Although this complicates the interpretation of the test results, it is a realistic test set.

TABLE 5. Comparison of recognition rates

	Percentage Correct	Number of Tests
16-bit 8kHz PCM	90%	333
LPC	73%	320
CWELP	76%	341
CELP	85%	338

Table 5 shows that of the three vocoders CELP performed the best. CWELP appears to have performed slightly better than LPC. However, the difference is not significant enough to make a judgement as to which technique provides superior speech quality.

6.3.1 Common Mistakes

Table 6 shows the most common misidentifications and their percentage contribution to the total number of misidentifications. The z/v misidentification was augmented because the ISOLET database is American, whereas the listeners were Canadian -- they expected "zed," not "zee".

TABLE 6. Common misidentifications

	Percentage of Total Misidentifications
d, b, and e	16%
s and f	10%
z and v	8%
n and m	6%
Other	60%

6.4 Informal Sentence Tests

Two females and two males recorded the sentence "The quick brown fox jumped over the lazy dog's back." When randomly presented with the LPC and CWELP synthesized sentences, listeners consistently chose the CWELP sentences as being of superior sound quality. CWELP-synthesized speech tends to be consistently noisy. LPC-synthesized speech is mostly clear, but an occasional pure tone or misplaced unvoiced sound will occur, distracting the listener. The listener can become used to the consistent noise of CWELP, but is more confused by the unpredictable LPC noise.

6.5 Non-human Speech Coding

In order to test the ability of the vocoder to reproduce sounds with extraordinarily complex excitation, bat calls and whale calls were compressed. The original bandwidths of 80kHz and 10kHz respectively were reduced to 4kHz by simply reinterpreting the sample rate for the calls. Whereas the CWELP vocoder reproduced the bat calls and whale calls reasonably accurately (to the human ear), the LPC vocoder could not reproduce the bat calls at all and generated extremely mechanical-sounding whale call reproductions.

This experiment illustrates the ability of the CWELP vocoder to correctly compress sounds which have more complicated forms of excitation.

Chapter 7: Conclusions

The CWELP vocoder performs as expected; it is a compromise of speech quality and computational complexity between the LPC vocoder and the CELP vocoder as shown in Figure 35.

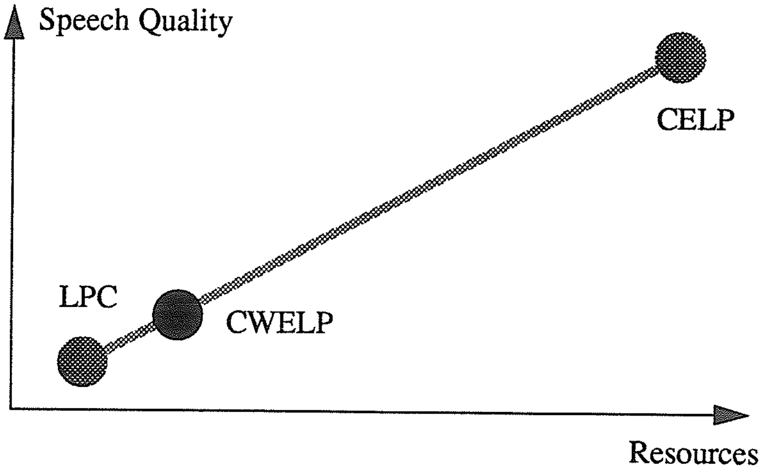


FIGURE 35. Qualitative comparison between LPC, CWELP, and CELP. "Resources" refer to computational complexity and bit rate.

Whether or not the CWELP compression scheme is of any use in the practical market can be determined only by exhaustive experimentation. Popular coding methods such as ADPCM, LPC, and CELP have been optimized and well-tested by thousands of researchers receiving support from the telecommunications industry.

Low-level improvements such as delta-coding the vocal tract parameters and the wavelet coefficients would further reduce the CWELP bit rate.

Other DWT bases may lead to superior sound quality. Indeed, another short-term excitation coding scheme altogether may outperform the DWT method. Another alternative for coding the short-term excitation is to combine impulse coding and DWT coding.

Further improvement could also be obtained by applying a weighting filter to the true excitation before coding it. Such a weighting filter should emphasize excitation components that are *critical to the human ear after the vocal tract filter is applied.*

Regardless, at its simplest, the CWELP vocoder performs well.

Appendix A: List of English Phonemes

TABLE 7. Classification of English phonemes along with examples

Classification			Orthographic and Phonetic Representations		Example
Vowels		Front	i	IY	beet
			I	I	bit
			e	E	bet
			æ	AE	bat
		Mid	a	A	hot
			-	ER	bird
			Λ	UH	but
			o	OW	bought
		Back	u	OO	boot
			U	U	foot
			O	O	bow
Diphthongs			aI	AI	bay
			oI	OI	boat
			aU	AU	buy
			eI	EI	how
			oU	oU	hey
			ju	JU	you
Semivowels		Liquids	w	W	walk
			l	L	lock
		Glides	r	R	rock
			y	Y	yacht
Consonants	Nasals		m	M	rum
			n	N	run
			ŋ	NG	rung
	Stops	Voiced	b	B	bun
			d	D	done
			g	G	gun
		Unvoiced	p	P	pun
			t	T	ton
			k	K	come
	Fricatives	Voiced	v	V	very
			∂	TH	this
			z	Z	zoo
			zh	ZH	Jaque
		Unvoiced	f	F	ferry
			θ	THE	thistle
			s	S	sue
			sh	SH	shack
	Affricates		tsh	TSH	witch
			j	DZH	jack
	Whisper		h	H	hot

Appendix B: Using Daubechies' Wavelet Filter to Generate the DWT

Consider the following linear transform matrix:

$$\begin{bmatrix} c_0 & c_1 & c_2 & c_3 & & & & & & \\ & & c_0 & c_1 & c_2 & c_3 & & & & \\ & & & & & \cdots & & & & \\ & & & & & & c_0 & c_1 & c_2 & c_3 \\ c_2 & c_3 & & & & & & & c_0 & c_1 \\ c_3 & -c_2 & c_1 & -c_0 & & & & & & \\ & & c_3 & -c_2 & c_1 & -c_0 & & & & \\ & & & & & \cdots & & & & \\ & & & & & & c_3 & -c_2 & c_1 & -c_0 \\ c_1 & -c_0 & & & & & & & c_3 & -c_2 \end{bmatrix} \qquad (22)$$

This matrix has the form of two convolution transforms with every other sample skipped. The net effect is that two types of filtering operations occur. If $\{c_0, c_1, c_2, c_3\}$ form a smoothing-type filter (like a lowpass filter), then $\{c_3, -c_2, c_1, -c_0\}$ form a detail-type filter (like a highpass filter).

This transform will be used to construct the wavelet transform, so it is desirable that it be orthonormal (each row and column has a norm of unity, and the inner product of any two rows or columns is zero). The normality condition requires

$$c_0^2 + c_1^2 + c_2^2 + c_3^2 = 1 \qquad (23)$$

Most of the rows and columns in Equation 22 are explicitly orthogonal, but another requirement results from the orthogonality condition:

$$c_0 c_2 + c_1 c_3 = 0 \qquad (24)$$

Since there are four coefficients and only two restrictive equations above, the system still has two degrees of freedom. Daubechies' filter coefficients are constructed by requiring the detail filter to have a certain number of vanishing moments. Two degrees of freedom can be removed by requiring the first two moments to be vanishing ("approximation condition of order 2"):

$$c_3 - c_2 + c_1 - c_0 = 0 \tag{25}$$

$$0c_3 - 1c_2 + 2c_1 - 3c_0 = 0 \tag{26}$$

Equations 23 to 26 give rise to the following Daubechies filter coefficients:

$$\begin{aligned} c_0 &= (1+\sqrt{3})/4\sqrt{2} & \qquad c_1 &= (3+\sqrt{3})/4\sqrt{2} \\ c_2 &= (3-\sqrt{3})/4\sqrt{2} & \qquad c_3 &= (1-\sqrt{3})/4\sqrt{2} \end{aligned} \tag{27}$$

These are the so-called DAUB4 filter coefficients. Higher order filters, such as the DAUB20 filter, are constructed in a similar manner, using higher order approximation conditions (the number of vanishing moments is greater).

The discrete wavelet transform (DWT) is obtained by hierarchically applying the wavelet filter. First, the filter is used to extract the finest detail via the detail filter. The remainder, generated by the smooth filter consists of a lower-sampled version of the smooth part of the input vector. When the wavelet filter is applied again to this smoothed vector, the detail filter extracts lower-frequency detail, while the smooth filter generates a smooth-smooth vector. In this way, the DWT extracts high-frequency detail all the way down to low-frequency detail. The results of repetitive application of the wavelet filter are shown in Equation 28 on page 55. Notice that if the input vector is of size *N*, the first wavelet filter is of size *N* x *N*, the second filter is of size *N*/2 x *N*/2, the third filter is of size *N*/4 x *N*/4, and so on. The raised roman index in Equation 28 refers to the number of filters used to obtain the coefficient. The output of the system consists of a hierarchy of detail coefficients, where the higher raised indices refer to courser detail. Also, two smooth coefficients are left (obviously, all of the smooth coefficients cannot be filtered away). These are called the "mother-function coefficients".

Since each filter coefficient is calculated using a fixed number of multiplications (four in the DAUB4 case), the total number of multiplications is of order *N* + *N*/2 + *N*/4 + *N*/8 + ... Consequently, the DWT is of order *N*.

Higher-order DWTs are obtained by satisfying the orthonormality and vanishing moment conditions (analogous to Equations 23 to 26) to obtain the wavelet filter. Again, the filter is applied hierarchically to perform the DWT.

For information on the theory and application of the DWT see [16] and [7].

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \\ x_9 \\ x_{10} \\ x_{11} \\ x_{12} \\ x_{13} \\ x_{14} \\ x_{15} \\ x_{16} \end{bmatrix} \rightarrow \begin{bmatrix} s_1^{(i)} \\ s_2^{(i)} \\ s_3^{(i)} \\ s_4^{(i)} \\ s_5^{(i)} \\ s_6^{(i)} \\ s_7^{(i)} \\ s_8^{(i)} \\ d_1^{(i)} \\ d_2^{(i)} \\ d_3^{(i)} \\ d_4^{(i)} \\ d_5^{(i)} \\ d_6^{(i)} \\ d_7^{(i)} \\ d_8^{(i)} \end{bmatrix} \rightarrow \begin{bmatrix} s_1^{(ii)} \\ s_2^{(ii)} \\ s_3^{(ii)} \\ s_4^{(ii)} \\ d_1^{(ii)} \\ d_2^{(ii)} \\ d_3^{(ii)} \\ d_4^{(ii)} \\ d_1^{(i)} \\ d_2^{(i)} \\ d_3^{(i)} \\ d_4^{(i)} \\ d_5^{(i)} \\ d_6^{(i)} \\ d_7^{(i)} \\ d_8^{(i)} \end{bmatrix} \rightarrow \begin{bmatrix} s_1^{(iii)} \\ s_2^{(iii)} \\ d_1^{(iii)} \\ d_2^{(iii)} \\ d_1^{(ii)} \\ d_2^{(ii)} \\ d_3^{(ii)} \\ d_4^{(ii)} \\ d_1^{(i)} \\ d_2^{(i)} \\ d_3^{(i)} \\ d_4^{(i)} \\ d_5^{(i)} \\ d_6^{(i)} \\ d_7^{(i)} \\ d_8^{(i)} \end{bmatrix} \tag{28}$$

Appendix C: Mathematics of Classical Linear Prediction

Given samples of data, *x(k)*, the goal of classical linear prediction is to determine a set of coefficients α_j, $j = 1, ..., p$ (where *p* is the predictor order) such that the series in Equation 29 is a good approximation to the data.

$$y(k) = \sum_{j=1}^{p} \alpha_j x(k-j) \tag{29}$$

The Euclidean metric is used to define an approximation error as

$$E = \sum_k [x(k) - y(k)]^2 = \sum_k \left[x(k) - \sum_{j=1}^{p} \alpha_j x(k-j) \right]^2 \tag{30}$$

where the sum should range over the portion of the data for which most accurate linear prediction is desired. **The formal definition of classical linear prediction** allows for any appropriately large range for *k* by making the following qualification: the data series is assumed to be stationary. This means that the local statistics of the data do not change with index *k*; in other words, the local statistics are the same as the global statistics. If we are interested in obtaining optimal linear prediction for **a subset** of the data series, then it makes more sense to expect that the data is nonstationary, but restrict the range of summation so that optimal coefficients are obtained for the subset.

The error in Equation 30 is minimized by setting $\partial E / \partial \alpha_i = 0, \forall i$:

$$\frac{\partial E}{\partial \alpha_i} = -\sum_k 2\,[x(k) - y(k)]\,\frac{\partial y(k)}{\partial \alpha_i} = -\sum_k 2 \left[x(k) - \sum_{j=1}^{p} \alpha_j x(k-j) \right] x(k-i) = 0 \tag{31}$$

Equation 31 simplifies to

$$\sum_{j=1}^{p} \alpha_j \left[\sum_k x(k-j) x(k-i) \right] = \sum_k x(k) x(k-i) \tag{32}$$

The following definition separates out the statistical constants from this equation:

$$\phi(i,j) = \sum_{k} x(k-j)x(k-i) \tag{33}$$

These constants can be calculated from the data series. Equation 32 becomes

$$\sum_{j=1}^{p} \alpha_j \phi(i,j) = \phi(i,0) \tag{34}$$

which is p equations in p unknowns. This equation may be solved using standard matrix techniques once the statistical constants are calculated. However, by appropriately choosing the index k over which the statistics are determined, various specialized techniques exist [1]. The linear predictive techniques used in this thesis were developed by Burg and are described in detail in [1] and [7].

Appendix D: The ISOLET Spoken Letter Database

This database was compiled by researchers at the Oregon Graduate Institute of Science and Technology [17]. The names and address of the authors is

Ron Cole, Yeshwant Muthusamy, Mark Fanty
Department of Computer Science and Engineering
Oregon Graduate Institute of Science and Technology
19600 N.W. Von Neumann Drive
Beaverton, OR 97006
E-mail: cole@cse.ogi.edu

ISOLET consists of 7800 letters spoken in isolation by 150 speakers. Each speaker uttered each letter of the alphabet twice. Data from all age groups was collected across Canada and the United States.

Table 8 lists the speakers whose utterances were used in Chapter 6.

TABLE 8. List of speaker's used in this thesis.

ISOLET I.D.	Sex	Age	Location
fcmc0	female	38	Oregon
fet0	female	60	Florida
fjw0	female	38	Oregon
fkh0	female	42	Montana
frw0	female	22	California
mjc1	male	17	Oregon
mjp0	male	41	New York
msa0	male	36	Oregon
mteb0	male	32	Washington
mwr0	male	58	Oregon

Data was recorded in a 15' by 15' tiled-floor room with standard office wall board and drop ceiling. The analog signal from a Sennheiser HMD 224 noise-cancelling microphone was lowpass filtered at 7.6 kHz and then sampled at 16 kHz.

Since the compression schemes used in this paper required data sampled at 8 kHz, the ISOLET data were digitally lowpass filtered at 3.8 kHz and then decimated by two.

The signal to noise ratio was calculated to be 31.5 dB with a standard deviation of 5.6 dB.

References

1. Rabiner, L.R. and Schafer, R.W., "Digital Processing of Speech Signals", Prentice-Hall Inc., New Jersey, USA, 1978.

2. Campbell, J.P. Jr., Tremain, T.E., and Welch, V.C., "The DOD 4.8 KBPS Standard (Proposed Federal Standard 1016)," in Atal, Bishnu S., Cuperman, V., Gersho, A., *Advances in Speech Coding*, Kluwer Academic Publishers, Massachusetts, 1991, p121.

3. Pickles, James O., "An Introduction to the Physiology of Hearing," Academic Press Inc., Toronto, 1988.

4. Chui, Charles K., "An Introduction to WAVELETS," Academic Press Inc., Toronto, 1992.

5. Daubechies, Ingrid, "The Wavelet Transform, Time-Frequency Localization and Signal Analysis," *IEEE Transactions on Information Theory*, vol. 36, no. 5, September, 1990, p961-1005.

6. Oppenheim, A.V., "Speech Spectrograms Using the Fast Fourier Transform," *IEEE Spectrum*, vol. 7, August, 1970, p57-62.

7. Press, W.H., Teukolsky, S.A., Vetterling, W.T., Flannery, B.P., "Numerical Recipes in C, The Art of Scientific Computing, Second Edition," Cambridge University Press, Cambridge MA, 1988, p591-606.

8. "Transmission Systems for Communication," Bell Telephone Laboratories, p73.

9. Paez, M.D., and Glisson, T.H., "Minimum Mean Squared-Error Quantization in Speech," *IEEE Trans. Comm.*, vol. Com-20, p225-230, April, 1972.

10. Smith, B., "Instantaneous Companding of Quantized Signals," *Bell System Tech. Journal*, vol. 36, no. 3, p653-709, May, 1957.

11. Schafer, R.W., and Rabiner, L.R., "Design and Simulation of a Speech Analysis-Synthesis System Based on Short-Time Fourier Analysis," *IEEE Trans. Audio and Electroacoustics*, vol. AU-21, no. 3, p165-174, June, 1973.

12. Wickerhauser, M.V., "Acoustic Signal Compression with Wavelet Packets," in Chui, C.K., *Wavelets - A Tutorial in Theory and Applications*, Academic Press Inc., Toronto, 1992, p679-700.

13. Atal, B.S., Hanauer, S.L., "Speech analysis and Synthesis by Linear Prediction of the Speech Wave," *J. Acoust. Soc. Am.*, vol. 50, p637-655, 1971.

14. Atal, Bishnu S., Cuperman, V., Gersho, A., "Advances in Speech Coding," Kluwer Academic Publishers, Norwell, MA, 1991.

15. Welch, V., Tremain, T., and Campbell, J., "A Comparison of U.S. Government Standard Voice Coders," *Proceedings of the IEEE Military Communications Conference (MILCOM)*, 1989, p269-273.

16. Daubechies, I., *Communications on Pure and Applied Mathematics*, vol. 41, 1988, p909-996.

17. Cole, R., Muthusamy, Y. Fanty, M., "The ISOLET Spoken Letter Database", *Technical Report No. CSE 90-004*, Department of Computer Science and Engineering, Oregon Graduate Institute of Science and Technology, March, 1990.