Identification of Birdsong using Artificial Neural Computing.

by Alex L. McIlraith

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

© Alex L. McIlraith 1996

National Library
of Canada

Acquisitions and
Bibliographic Services Branch

395 Wellington Street
Ottawa, Ontario
K1A 0N4

Bibliothèque nationale
du Canada

Direction des acquisitions et
des services bibliographiques

395, rue Wellington
Ottawa (Ontario)
K1A 0N4

Your file Votre référence

Our file Notre référence

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

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

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

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

ISBN 0-612-16207-9

Canada

Name ___

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

Electronics & Electrical Engineering

SUBJECT TERM

0544

SUBJECT CODE

UMI

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
- Médiévale 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-être 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 pâturages 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 hôpitaux 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
- Généralités 0485
- Biochimie 487
- Chimie agricole 0749
- Chimie analytique 0486
- Chimie minérale 0488
- Chimie nucléaire 0738
- Chimie organique 0490
- Chimie pharmaceutique 0491
- Physique 0494
- PolymÈres 0495
- Radiation 0754

Mathématiques 0405
Physique
- Généralités 0605
- Acoustique 0986
- Astronomie et astrophysique 0606
- Électronique 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 thermodynamique 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 systèmes 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

THE UNIVERSITY OF MANITOBA

FACULTY OF GRADUATE STUDIES

COPYRIGHT PERMISSION

IDENTIFICATION OF BIRDSONG USING ARTIFICIAL NEURAL COMPUTING

BY

ALEX L. MCILRAITH

A Thesis/Practicum 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

Alex L. McIlraith © 1996

Permission has been granted to the LIBRARY OF THE UNIVERSITY OF MANITOBA to lend or sell copies of this thesis/practicum, to the NATIONAL LIBRARY OF CANADA to microfilm this thesis/practicum and to lend or sell copies of the film, and to UNIVERSITY MICROFILMS INC. to publish an abstract of this thesis/practicum..

This reproduction or copy of this thesis has been made available by authority of the copyright owner solely for the purpose of private study and research, and may only be reproduced and copied as permitted by copyright laws or with express written authorization from the copyright owner.

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.

Alex L. McIlraith 1996

CONTENTS

1. LITERATURE REVIEW 1

1.1 SOUND PROCESSING IN VERTEBRATES 1
1.1.1 Defining 'song' 2
1.1.2 Ecological significance of bird sound 2
1.1.3 Mechanics of human speech 5
1.1.4 Mechanics of bird sound 9
1.1.5 Neural processing of sound 13
1.2 PATTERN RECOGNITION 14
1.2.1 Pre-processing 14
1.2.2 Statistical methods 23
1.2.3 Artificial neural network (ANN) methods 32
1.2.4 Comparison of methods 40
1.3 ANALYSIS OF BIRD SOUND 43
1.3.1 Preparation for analysis 44
1.3.2 Features for discrimination 44
1.3.3 Variability in song 48
1.3.4 Statistical analysis 51
1.4 JUSTIFICATION AND GOALS OF RESEARCH 52

2. SPECIES IDENTIFICATION - METHOD ONE 53

2.1 INTRODUCTION 53
2.2 METHODS 54
2.2.1 Sampling 54
2.2.2 Preprocessing 54
2.2.3 Identification 55
2.3 RESULTS 56
2.4 DISCUSSION 58

3. SPECIES IDENTIFICATION - METHOD TWO 60

3.1 INTRODUCTION 60
3.2 METHODS 61
3.2.1 Sampling 61
3.2.2 Pre-processing 61
3.2.3 Statistical analysis 62
3.2.4 Neural network analysis 63
3.3 RESULTS 63
3.3.1 Statistical analysis 63
3.3.2 Neural network analysis 69
3.4 DISCUSSION 72
3.4.1 Exploration of the data 72
3.4.2 Identification 74

4. CONCLUSION 76

5. LITERATURE CITED 77

6. APPENDICES 85

6.1 ENVELOPE DETERMINATION FROM LPC WITH MATHCAD 86
6.2 PRINCIPAL COMPONENTS ANALYSIS WITH MATHCAD 87
6.3 TUKEY BOX PLOTS OF FEATURE VARIABLES 90

ABSTRACT

Birds have long caught the interest of people. Recently birds have become a focus of attention as indicators of environmental quality, and as part of the growing hobby of bird-watching. Most birds make sounds, many of which are distinctive enough that one can identify a species by its sound. In fact, population censuses are often performed by identifying birds by ear. This has raised recent interest in finding a means for the automatic identification of birds by their sounds.

Extensive literature in the areas of human speech, speech recognition, signal processing and bird song provides a good starting point for automatic identification. Like speech, bird songs have characteristic temporal and spectral qualities. Techniques such as linear predictive coding (LPC) and power spectral density (PSD) analysis have been used successfully in the past for human speech recognition, and it makes sense that they would be applicable in recognizing bird sounds. Spectral information combined with temporal cues provides useful input for pattern recognition methods including neural networks or discriminant analysis.

In this study, 133 recorded bird songs of six species were sampled (11.025kHz, 8bits/sample) and digitally pre-processed. From sampled data, spectral and temporal measurements were made. Two different approaches were taken in pre-processing and subsequent identification.

The first approach was to extract LPC coefficients for 512 or 2048 sample windows of each song, and to present the window information to a back-propagation neural network. Information for the two window sizes combined was also presented. In addition, the network was also provided with temporal information about the length of the song associated with each window. Seventy-five percent of the data was used in training, and the remaining 25% used for testing the network's performance. The number of frames placed into each species class was counted for test songs; the bird was identified as belonging to the class which was represented by the largest number of frames. Ten training trials (1500 epochs) of each network configuration (10 or 19 inputs, 12 hidden nodes and 6 outputs) were used, with the overall identification accuracy ranging from 91% to 93%.

The second approach involved explicitly measuring temporal characteristics such as the mean, standard deviation and number of sounds (elements) within a song. In addition, PSDs were calculated for each sound period and the mean and standard deviation of these were calculated within the song. Using stepwise discriminant analysis, a subset of these variables (8 out of 23) was extracted for use in subsequent analyses. ANOVA, principal components analysis and canonical discriminant analysis were used to examine patterns within the subset.

Identification of records in the test sets involved the use of a back-propagation neural network and quadratic discriminant analysis. Experimentation suggested that a network with eight inputs, six hidden units and six outputs trained over 200 epochs was appropriate. Accuracy was 82%. However, with discriminant analysis, this method achieved an accuracy of 93.3%. Further research would be required to determine whether or not additional tuning of the neural network architecture and parameters could improve its performance relative to discriminant analysis.

Given the result, the second approach, with discriminant analysis, appears to be the best choice for two reasons. First, it required less computational time since a smaller network with fewer training epochs was required. Secondly, identification accuracy was essentially equivalent to that of the first approach.

ACKNOWLEDGEMENTS

First, I would like to thank my advisor Dr. Howard Card for his support, patience and scholarly tutelage. I thank my wife Sharon McIlraith for her continual encouragement and patience. Thank you to Brad Brown for getting me started on this road. Robert Berger and James Ehnes made many valuable suggestions and edited early drafts of the thesis. Charles Burchill critiqued preparations for the oral presentation. Thank you also to thesis committee members Dr. R.D. McLeod and Dr. S.E. Cosens for their comments and suggestions.

As with any undertaking, there are many people who contribute ideas or stimulate thought. Those students who participated in the Ganglion group and its seminars, the VLSI group, those with whom I took courses, and others all helped in one way or another to make this M.Sc. reality; Andy Brown, Bob Cielen, Derek Ross, Brendan Frey, SriGouri Kamarsu, Jeff Kopytko, Gord McGonigal, Dean McNeill, Tom Tessier, and Richard Wieler all made special contributions. Support from the departments of Electrical and Computer Engineering and Graduate Studies is gratefully acknowledged.

This research was supported by NSERC and a University of Manitoba Fellowship.

This thesis is dedicated to Dick Convery.

Faery Bird's Song

Shed no tear - O, shed no tear!
The flower will bloom another year.
Weep no more! O! weep no more!
Young buds sleep in the root's white core.
Dry your eyes! O! dry your eyes,
For I was taught in Paradise
To ease my breast of melodies -
Shed no tear.

Overhead! look overhead!
'Mong the blossoms white and red -
Look up, look up. I flutter now
On this flush pomegranate bough.
See me! 'tis this silvery bill
Ever cures the good man's ill.
Shed no tear! O shed no tear!
The flower will bloom another year.
Adieu, adieu - I fly, adieu,
I vanish in the heaven's blue -
Adieu, adieu!

-John Keats

LIST OF FIGURES

FIG. 1.1 SCHEMATIC OF THE HUMAN VOCAL TRACT 5
FIG. 1.2 WINDOWED FOURIER TRANSFORM OF THE VOWEL SOUND IN 'BAAA' 7
FIG. 1.3 THE HUMAN EAR 8
FIG. 1.4 SCHEMATIC OF A FRONTAL SECTION THROUGH A VOCAL TRACT 10
FIG. 1.5 EFFECT OF INCREASING M ON THE PSD OF A RANDOM SIGNAL 19
FIG. 1.6 LPC ANALYSIS OF A WINDOW FROM THE VOWEL SOUND IN 'BAAA' 22
FIG. 1.7 A PROCESS FOR CHOOSING RELEVANT ANALYSIS METHODS 24
FIG. 1.8 DENDROGRAM FOR NEAREST NEIGHBOR HIERARCHICAL CLUSTERING 32
FIG. 1.9 TIME DOMAIN REPRESENTATION OF A FOX SPARROW SONG 47
FIG. 1.10 FREQUENCY DOMAIN REPRESENTATION OF A FOX SPARROW SONG 48
FIG. 1.11 SPECTROGRAMS OF THREE FOX SPARROW SONGS 50
FIG. 2.1 SCHEMATIC EXAMPLE OF PRE-PROCESSING 55
FIG. 2.2 LEARNING CURVES FOR A BACK-PROPAGATION NETWORK 57
FIG. 2.3 SPECIES RECOGNITION PERFORMANCE FOR THREE NEURAL NETWORKS 57
FIG. 3.1 THE FIRST THREE PRINCIPAL COMPONENTS 67
FIG. 3.2 THE FIRST TWO CANONICAL DISCRIMINANTS 69
FIG. 3.3 IDENTIFICATION ACCURACY WITH DISCRIMINANT ANALYSIS 69
FIG. 3.4 EFFECT OF NUMBER OF HIDDEN UNITS, AND LENGTH OF TRAINING PERIOD ON GENERALIZATION 70
FIG. 3.5 IDENTIFICATION ACCURACY WITH A BACK-PROPAGATION NEURAL NETWORK 71
FIG. 3.6 OVERVIEW OF ANALYSIS PROCEDURE 72

LIST OF TABLES

TABLE 1.1 DISTANCE MATRIX FOR FIVE RECORDS 31
TABLE 1.2 NEAREST-NEIGHBOR GROUPS 32
TABLE 1.3 COMMON TRANSFER FUNCTIONS FOR ARTIFICIAL NEURONS 33
TABLE 1.4 A COMPARISON OF MULTIVARIATE STATISTICS AND ARTIFICIAL NEURAL NETWORKS 41
TABLE 1.5 IMPORTANCE OF CERTAIN SOUND CHARACTERISTICS FOR SPECIES RECOGNITION IN BIRDS 45
TABLE 2.1 EXAMPLE OF NEURAL NETWORK OUTPUT 56
TABLE 2.2 LENGTH OF SONGS 58
TABLE 3.1 STEPWISE DISCRIMINANT ANALYSIS OF TEMPORAL AND SPECTRAL VARIABLES 64
TABLE 3.2 CONTRIBUTION OF VARIABLES TO THE DISCRIMINANT MODEL 64
TABLE 3.3 ANALYSIS OF VARIANCE SUMMARY 65
TABLE 3.4 PCA, USING CORRELATION 66
TABLE 3.5 EIGENVECTORS FOR THE FIRST THREE PCA AXES 66
TABLE 3.6 CDA FOR THE EIGHT VARIABLE DATA SET 68
TABLE 3.7 EIGENVECTORS FOR THE TWO CDA AXES 68

1. LITERATURE REVIEW

Birds have been of interest to people since history began to be recorded. Both the social and ecological importance of birds is reflected in the laws we have instituted to protect them. In Canada and the United States, migratory bird legislation makes it illegal to kill or collect migratory species without a permit (Canadian Wildlife Service, 1980). In addition to ongoing bird research (ornithology), a growing number of people are taking up bird-watching as a hobby. One can easily argue that birds are considered an important part of many people's lives.

Although some biologists are interested in bird sounds themselves, others consider sounds a valuable tool for identification. In ornithological research, biologists may have interest in the variety of sounds made by individuals or groups of individuals within a species, or in the variety of sounds made by different species in an area. Skill in identifying sounds is often helpful for assessing the density of breeding birds. Such information can be used to assess long-term trends in numbers of one or more species, or to compare numbers between different locations and/or times (Ralph et al., 1993).

Biologists are often called upon to predict or assess the impact of human development on plants and animals. In the process of executing such an assessment, the biologist may have to identify and count birds in a site. As many of the birds in an area may be heard and not seen, it is often quicker and easier to rely on sounds. Thus the biologist must study the sounds of birds for that area and be able to identify them by sound alone. Since this can be a difficult task, it can be helpful to record unknown sounds so that they can be identified later.

Recently, interest has arisen in the possibility of building a device that could not only record but automatically identify birds (and perhaps other animals) by their sounds. This would have the advantage that it would not require an expert user. The suggestion has also been made that such technology could be used as a long term monitoring device (McKenna, 1995; Mittelstaedt, 1994). Other applications for such technology could be systems for education and corroboration of human accuracy.

In order to fully appreciate the problem of automatic species identification using sound, background in the areas of biology, computer engineering and statistics is beneficial. Biology provides information about why and how animals make sounds. Knowing this, in turn suggests some features of sounds that might be important in distinguishing species. Computer engineering provides the theory required to extract these features. Computer engineering and statistics supply analysis methods for automatic identification. This literature review provides a background and/or a review to the reader in these three areas.

1.1 Sound processing in vertebrates

The goal of this section is to give the reader an appreciation for the biology of bird sound. First, however, issues surrounding the use of the term 'song' are examined to avoid confusion. Next, it is useful to understand why birds use sound: its ecological

significance. After this, the mechanics and physiology associated with communication via sound in vertebrates becomes significant. Since the mechanics of vocalization and audition in humans are well known, they are discussed to provide a background for an examination of bird sound. Lastly, neurological research on owls and bats is introduced to help explain neural data processing used by vertebrates in audition.

1.1.1 Defining 'song'

The general term 'sound' is often used in preference to 'song' or even 'vocalization' when describing sounds made by birds in this manuscript. The main reason is that there are both theoretical and practical difficulties with these terms. The words sound and vocalization imply that the sound comes from the bird's mouth. In reality most bird sounds do originate in the vocal tract. However, birds such as the common snipe, ruffed grouse and woodpeckers make at least some of their sounds using specialized feathers, their wings, or by tapping on objects (Godfrey, 1979). In the first two species these sounds are used in mate attraction, which is one of the functions of song. Spector (1994) has examined the historical use of the term 'bird song' and concludes that there is little agreement on a definition. It is suggested that researchers should carefully define the term as it is used in their work.

Thorpe (1961) provides a practical framework for a definition. Song is generally restricted to a sub-group of the Passeriformes (the perching birds), called the oscines (song birds); this group was initially delineated on the basis of syringeal structure (the syrinx is the vocal organ in birds). In birds it is generally the male that sings in order to attract mates, and to define the boundaries of its territory. Calls and call notes, distinguished by their shorter duration, are produced by both sexes. They tend to be used to signal alarm or recognition. Although songs tend to be characteristic of the species generating them, call notes may not be. I propose the following working definition for song: song is a vocal sound that is generally produced by males in a breeding or territorial context; songs are generally more complex than call notes produced by the same species; unlike Thorpe's definition, this one is not restricted to oscines alone. This definition is in line with criteria listed by Spector (1994).

In this study, characteristic sounds were required in order to identify birds to species. For the sake of generality the term 'sound' is often used. Since many birds do produce distinctive songs, these are commonly what is used by biologists to identify them. All species analyzed in this study have songs which fit the working definition.

1.1.2 Ecological significance of bird sound

Biologists attempt to explain how and why organisms live, reproduce and interact with their environment. One fruitful consensus that has emerged is that one must always consider the evolutionary implications (past and present) of a given behavior or characteristic of an individual. T. Dobzhansky issued a now famous statement to this effect (Curtis, 1983):

"Nothing in biology makes sense except in the light of evolution."

Although we may describe behaviors in birds with a term like "it wants to sing", this is misleading because it is anthropomorphic (casting something in a human-like light). The more pragmatic approach is to try to explain song as a behavior that has past and future evolutionary benefits. The thought process can be demonstrated with a hypothetical example. Imagine that some males of a bird species are born with a genetic inability to sing. It is likely that these individuals will fail, on average, to find mates since singing is important in mate attraction. They are unlikely to contribute many of their genes to the next generation. In time, these non-singing birds become rare or are eliminated within the species.

In efforts to understand the significance of bird sound, some basic questions that one may ask are:

- Why do birds make sounds?

- What factors have influenced the course of sound evolution?

1.1.2.1 Purpose of vocalization

Bird song is important in attracting mates and repelling non-mates (Morse, 1980). Generally, only male birds sing. Males are generally more brightly colored. These males are advertising themselves as suitable mates for females. The female chooses her mate based on such characteristics as song, plumage, and the quality of the territory a particular male holds. Song also serves in defining and defending territories (Morse, 1980). In fact, males on adjacent territories will often sing one after the other in what is called countersinging (Farabaugh, 1982). Occasionally, both birds of a pair will sing. Such duetting has territorial and pair-bonding functions. Vocalizations, including call notes, are also used to communicate a variety of information related to states and activities of: distress, territorial defense, flight, feeding, courtship, mate feeding, nesting, flocking, roosting, aggression, begging (by nestlings) and alarm among others (Morse, 1980).

1.1.2.2 Selective pressures

It is unlikely that we will ever know what factors have influenced the course of evolution (selective pressures) and have molded the vocalizations of a given bird species. This however does not preclude the development of evolutionary theories about the origin of these sounds. Endler (1992), in a review, discusses the use of signals (auditory and otherwise) by animals for communication. He presents several conclusions that are relevant to our understanding of bird sound.

The systems animals use to generate and receive signals are the product of evolution. The direction of evolution in a given species is and has been affected by many factors including: the physics of the channel existing between sender and receiver, signaling behavior, habitat choice, and the nature of the signal-generating and signal-receiving

systems. The use of organs for several functions has imposed conflicting optimization constraints on these organs, and has resulted in compromise solutions. The successful use of a particular signal also depends on the nature of the habitat and channel conditions (e.g. the presence of noise or attenuation). It is also important to signal in a manner which does not attract predators. Even with such a complicated array of selection pressures, it is possible make some broad generalizations that apply to signaling with sound (Endler, 1992):

a) Use of frequencies below 2 kHz to reduce reverberation, attenuation and scattering by the environment in some habitats.

b) Use of frequencies above 0.5-1 kHz to reduce energy requirements and destructive interference.

c) Where reverberation is a problem (e.g. forests), rapid repetition of sounds are avoided.

d) Frequency modulation (FM) is preferable to amplitude modulation (AM), since it is less affected by fading and reverberation.

e) Redundancy (repetition) is added if the channel is noisy.

f) Higher frequencies are better if information is to be transmitted with time restrictions because high frequency receptors have faster response times and because fading is less destructive to the signal. Fading is especially prevalent under windy conditions.

g) Communications are restricted to bands that are not used by other species, and tuning is performed to receiver-specific signals. Note that the former could also be accomplished by segregating communication in time and/or space.

h) Use of signals that degrade quickly for short-range communication and ones that degrade slowly for longer-range communication.

i) Use of a simple alert signal to capture the receiver's attention before sending a message.

Although there are exceptions to these rules, they are helpful in selecting features useful for identifying species. It makes sense that different species sharing the same habitat might in some cases use signals differing in frequency (within the range of human speech) and pattern in order to reduce interference. FM modulation is likely to predominate, making volume relatively unimportant. Sound repetition rates should vary among species. Redundancy and repetition could vary with channel conditions.

Birds must be able to identify songs of their own species. In order to understand how they do this, one must understand the communication process as it occurs in vertebrates. The most detailed examination of communication in vertebrates has been the research on human speech.

1.1.3 Mechanics of human speech

1.1.3.1 Vocalization

In humans, voiced sounds originate from the 'vocal cords' of the larynx (Fig. 1.1). As a note, the larynx originally evolved as a structure to prevent foreign materials from entering the lungs (Lieberman, 1984). As Lieberman and Blumstein (1988) point out, these 'cords' are actually membranous folds; the proper term for them is 'vocal folds' (Sataloff ,1992). The gap between the vocal folds is the glottis. When vocal fold muscles initially contract, the tension on them stops the flow of exhaled air from the lungs. Air pressure builds and separates the folds. Soon after separation, suction from high speed air passing between the folds (Bernoulli force) and the continuing muscle tension draw the folds together again. This expels periodic impulses of air into the vocal tract. Glottal excitation generates harmonics throughout and beyond the 4kHz telephone speech bandwidth. The pitch, or fundamental frequency, of voiced speech is 80-300Hz in males, generally ranges to 500Hz in women and children, and can reach an extreme of 1.5kHz in some cases. The excitation of unvoiced speech is usually created by a turbulent flow of air out of the mouth and can be described as band limited white noise, where frequencies range into the ultrasonic region.

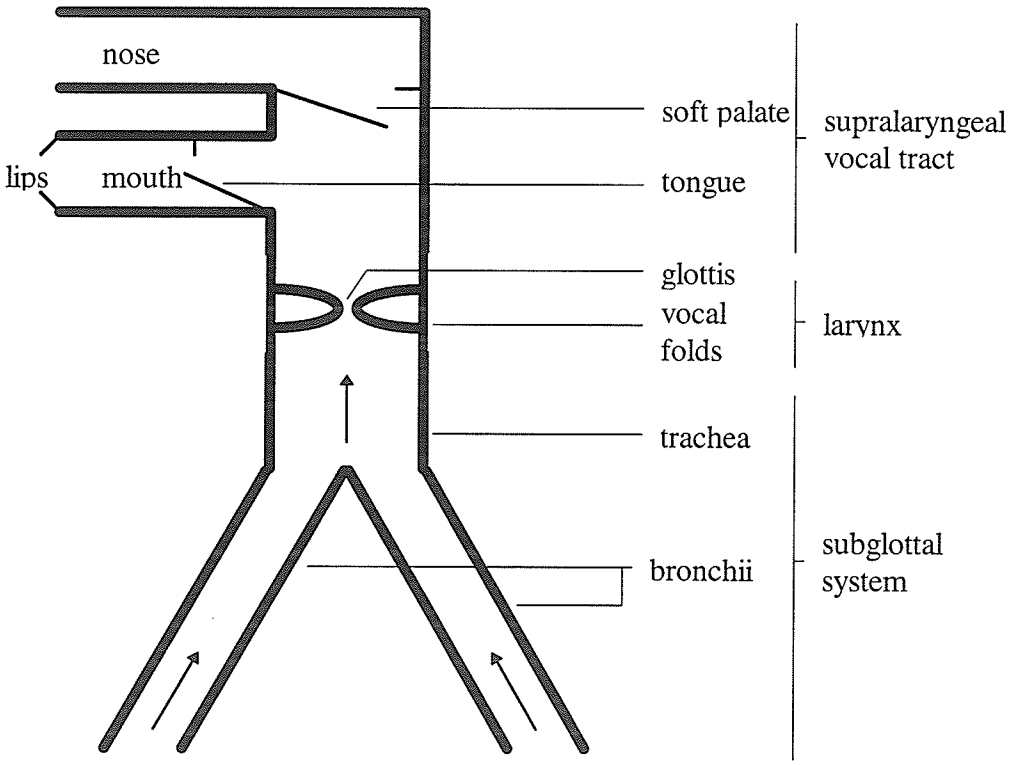


Fig. 1.1 Schematic of the human vocal tract; arrows indicate direction of air flow (after Lieberman and Blumstein, 1988).

Power input to the vocal system is provided by the lungs in the form of pressurized air. The volume of the sound is controlled by the force applied with the lungs. The fundamental frequency of voiced speech (called its pitch or glottal excitation) is controlled by varying air pressure and by muscles of the larynx. The region located above the glottis and larynx (the supralaryngeal vocal tract) includes the pharynx, oral cavity, nasal cavity, tongue, mouth and nose. The shape of these structures varies under muscular control, and modulates the frequency spectrum of the excitation. The supralaryngeal region forms a filter composed of a tube that is open at the exit end (mouth and /or nose).

In a simple open tube model, the cross-sectional area function of the tube determines the transfer function of the supralaryngeal region. One can approximate the continuous area function of the region as a number of concatenated tubes, with the resulting filter being an all-pole filter (see section 1.2.1.5). In speech terms, poles are equivalent to the damped resonances of the vocal tract filter. If more than one tube is involved, as is the case when opening of the soft palate (or velum) connects the nasal cavity to the supralaryngeal airway; this adds anti-resonances or zeros to the transfer function. Saying the letter 'n' results in such a situation. The maxima (poles) of the frequency envelope are termed formant frequencies, while the minima are called zeros. Although there may be many formants, speech remains intelligible if only those at the three lowest frequencies are retained. Fig. 1.2 indicates the frequency domain structure of a vowel, with three clearly visible formants. Varying the pitch does not necessarily affect the identity of the spoken vowel. Although women and children usually have significantly higher pitched speech than men, formant frequencies vary less.

A few simple exercises demonstrate the basic concepts of speech production.

1. Voiced speech involves movement of the vocal folds. Place your hand on your larynx (near the v-shaped spot on your neck known as Adam's apple), and hum a note. The vibration you feel comes directly from your vocal folds. It has a low frequency (its pitch) and many harmonics.

2. Open your mouth while still humming. You are now emitting a voiced sound. If you press on your chest or exhale more forcefully, the sound volume and/or frequency changes.

3. Now widen and narrow your the space between you lips, or move your tongue back and forth. Your sound seems to change in pitch. What is actually happening, is that the vocal folds are vibrating at a constant pitch, but the shape of your supralaryngeal vocal tract is emphasizing different harmonics.

4. If you say the word 'hello' normally, you are using voiced speech. If you whisper the word, you are using unvoiced speech. Turbulence in the glottis generates white noise as the excitation. It is filtered in the supralaryngeal region as it is during voiced speech.

5. If you say the letter 's', you will notice that the sound begins as a voiced consonant, but ends with an unvoiced hiss. The source of the unvoiced hiss is turbulent air flow through

the tongue and lips. Whispering the letter demonstrates a shift in location of the unvoiced excitation from the glottis to the mouth.

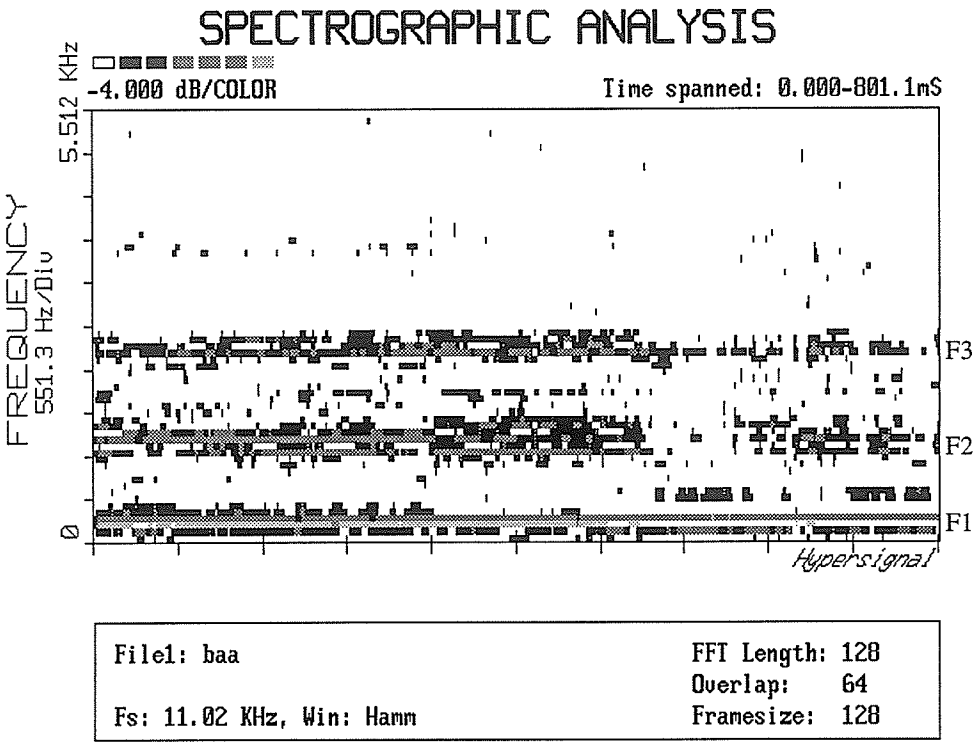


Fig. 1.2 Windowed Fourier transform of the vowel sound in 'baaa', showing three distinct formants (F1, F2, F3). Grey-scale boxes indicate the magnitude of the Fourier coefficients on a frequency-time plane (0-5512Hz, 0-801.1msec) in 4dB steps.

We now understand, in basic terms, what is meant by the source-filter theory of speech production. The source of excitation is either unvoiced or voiced. If it is voiced the fundamental pitch results from vibration of the vocal folds. The sounds are filtered by changing the shape of supralaryngeal structures (Lieberman and Blumstein, 1988). This simplified view of speech production leads the way toward techniques for electronic analysis and synthesis of speech. One such technique, linear predictive coding (LPC), is described later.

1.1.3.2 Audition

Sound signals are extensively processed before the signal reaches the brain . The face, pinnae, and auditory canal all filter the sounds we hear. For example, a given sound

arriving from an angle above the horizontal sounds different from one emitted from an angle below horizontal (Poon and Brugge, 1993). Sounds are coupled from the air to the liquid within the cochlea by the tympanum (ear drum), ossicles (bones) of the middle ear and oval window of the cochlea. (Fig. 1.3) These structures effect a mechanical transformer that matches the impedance of the air to that of the cochlear fluid (Whitfield, 1994). Vibrational energy is transmitted through the cochlear fluids and membranes (including the basilar membrane) to the round window located just below the oval window.

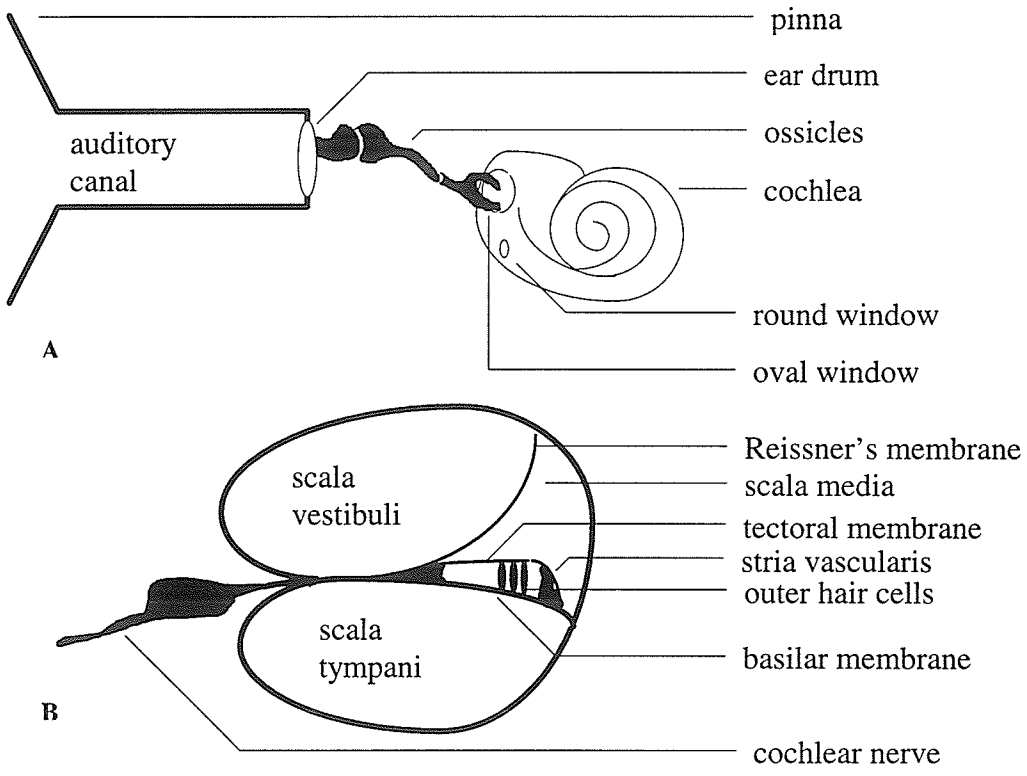


Fig. 1.3 The human ear (a), and a cross section of the cochlea (b) (after Eckert and Randall, 1983).

The static pressure outside the oval window is maintained at ambient levels by the Eustachian tube and respiratory tract. Static pressure changes and long wavelength sounds (20-30Hz) result in movement of fluid through the scala vestibuli and scala tympani; the wave thus travels the length of the cochlea from the oval window and back again to the round window (Eckert and Randal, 1983). As wavelengths become progressively shorter, waves still travel the length of the cochlea, but their maximum amplitude is localized nearer to the oval window.

The outer hair cells originate in the basilar membrane and terminate in the tectoral membrane above it. As the basilar membrane moves, shearing force is applied to the cells. A potassium ion pump in the stria vascularis generates about +80mV relative to the fluid of the scala tympani, while the resting potential of hair cells is about -60mV, for a total potential difference of 140mV. H. Davis (in Eckert and Randall, 1983) suggested that deflection of hair cell cilia varies the resistance with which the 140mV potential is applied to the hair cell. Upward deflection causes increased depolarization and hence neurotransmitter secretion, while downward motion reduces stimulation.

Neurotransmitter release modulates the firing rate of spontaneously firing neurons which carry the signal to the spiral ganglion and then the brain via the auditory nerve. Although the electromechanics of the cochlea provide some spectral decomposition of incoming sounds, they do not explain the narrow frequency discrimination abilities of mammals. The cochlea performs a time-frequency transform which has been termed a cochlear transform, the coefficients of which overlap (are non-orthogonal). Whitfield (1984) thus suggests that a 'second filter' must exist to reduce overlap and narrow the frequency resolution.

Neurons of the auditory cortex are arranged in a tonopic map and have a 'best frequency' or 'characteristic frequency' to which they respond most strongly. Whitfield (1984) suggests that some form of local lateral inhibition could produce the fine frequency resolution observed at this level.

The brain thus receives sound information in the form of nerve discharge spikes. Near the input to the auditory cortex, some cells within the tonotopic map are stimulated. Their position within the map indicates frequency, while the intensity of stimulation indicates magnitude. As a last point, studies have shown that resolution decreases approximately logarithmically (actually the fit is best with the square root transform) with frequency (Whitfield, 1984). In fact this logarithmic perception of frequency differences is the basis for the arrangement of musical notes (Deutsch, 1992).

With a reasonable understanding of the mechanics involved in our own communication, it is possible to compare and contrast them with those of birds.

1.1.4 Mechanics of bird sound

1.1.4.1 Vocalization

Brackenbury (1982, 1989) provides a clear description of the sound production mechanisms operating in birds; most of the information in this section is from this reference unless otherwise indicated.

The respiratory system of birds differs substantially from the human system. In humans, a series of progressively narrower bronchial tubes terminate in small air sacs in the lungs; the sacs themselves are used for gas exchange (Curtis, 1983). The system in birds is composed of a few large air sacs and many bronchial tubes. The numerous fine

parabronchial tubes form the lung proper, and are used for blood-air gas exchange. The sacs pull and push air through the parabronchi. The result is a very efficient and light-weight respiratory system. The air sacs have a low-resistance connection to the primary bronchus, in addition to the indirect path through the lungs. The direct path allows high velocity air to be applied to the syrinx. In birds, the larynx presumably has the function of preventing solids and liquids from entering the respiratory tract; it is the syrinx, not the larynx, which produces sound in birds.

The syrinx (Fig. 1.4) is located where the two primary bronchi join the trachea, within the interclavicular air sac. This location is critical in its function. Two tympanic membranes are located medially (toward the center of the body) where the bronchi join. A set of muscles of varying complexity, depending on the phylogenetic group to which the bird belongs, controls the operation of the syrinx. They do so by varying the tension on each tympanic membrane, sometimes independently. Other muscles may actually push the lateral labia toward the medial tympanic membranes, creating increased Bernoulli forces on the membranes.

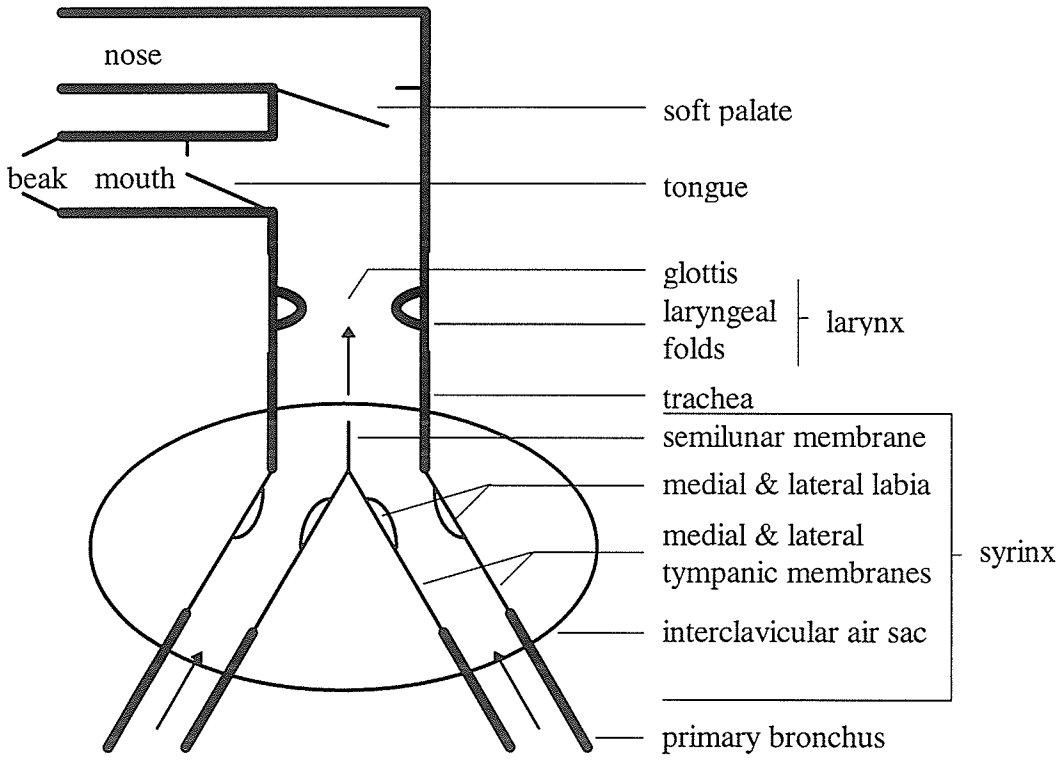


Fig. 1.4 Schematic of a frontal section through a vocal tract; arrows indicate direction of air flow (after Pettingill, 1985).

Sound is produced when a large amount of air pressure is applied to the syrinx. In order for the syringeal mechanism to operate, pressure in the interclavicular air sac (which surrounds the syrinx) must be approximately equal to the bronchial pressure. This is achieved by a direct connection between the sac and the bronchus. The forces which act to draw the tympanic membranes into the airway, balanced with controlled tension on the membranes results in oscillations as in the human vocal mechanism described above. Amplitude is controlled by varying the air flow rate. Two independent 'voiced' excitations can be generated through independent control of the right and left tympanic membranes. It appears that excitation of a bird's vocal tract is more sinusoidal than the impulse-like human excitation. This seems consistent with a lack of complete interruption of air flow.

In addition to the mechanisms just described, some birds appear to produce excitations as a result of air vortices in the syrinx. The Ringed Turtle Dove (Gaunt et al., 1982) for example produces pure tones or whistles, that are believed to be produced this way. It may be that many of the whistle-like sounds produced by birds are so generated. Largely unexplained is the ability of some species to generate three or four, harmonically unrelated, pitches. This coupled with the location of the syrinx at the base of the trachea in an air sac, its different mode of operation from that of laryngeal vocal folds, and the presence of auxiliary sources of sound in the vocal tract permit birds to produce sounds that humans cannot.

From an understanding of how a bird's vocal tract is excited, it is appropriate to focus on how the excitation is modulated and filtered. Birds modulate both the amplitude and frequency of their vocalizations, with modulation being classified as true (due to vibration of syringeal components) and spurious (due to characteristics of the vocal tract).

Spurious modulations are attributed to three factors. First, harmonic vibrations of the tympanic membranes modulate amplitude. Secondly, two tones of slightly different frequency and similar amplitude interact to create low frequency beat notes. Lastly, resonance within the trachea may affect the dominant frequencies emitted from the beak. It is well known that a tube open at both ends resonates differently from one that is closed at one end (Halliday and Resnick, 1981); resonant frequencies depend on the speed of sound in the tube air, the length of the tube, and whether or not the tube is open. If the trachea acts as a tube that is open at the syrinx and glottis, frequencies with ratios of 1:2:3... would be emphasized. If it acts as though the glottis was the only opening, frequencies with ratios of 1:3:5... would be emphasized. These frequencies are not necessarily harmonics of the fundamental excitation frequency. Supralaryngeal resonance properties could be varied by changing the length of the trachea.

Interaction between respiratory pressure and muscular control of syringeal components could produce true modulation of both amplitude and frequency. In fact pulsed contraction of respiratory muscles is used to produce warbled notes in the Ringed Turtle Dove (Gaunt et al., 1982). It is clear that birds use several methods of modulation in order to generate the complex range of sounds that has been recorded. Those interested in exploring the development of a detailed acoustical model for bird song production, should consult Fletcher (1988).

1.1.4.2 Audition

Sachs and O'Connell (In Lewis, 1983) have described audition from a neuro-physiological viewpoint. The cochlea of birds is uncoiled unlike that of mammals (Eckert and Randall, 1983) and is about 1/10 the length of the human cochlea (Dooling, 1982). Each fiber of the auditory nerve has its characteristic frequency (CF) and tuning curve. As in humans, tuning curves of birds overlap and become broader (in a linear manner) with increasing frequency even though their Q (CF / bandwidth at 10dB above threshold) increases. Unlike reptiles and amphibians, birds and mammals share the ability to hear frequencies across the audible spectrum. Interestingly, in both birds and mammals, the firing rates of auditory nerve fibers are phase-locked to the half-cycles of tones for frequencies below 6kHz.

As in humans, tonotopic maps have been found in avian (bird) brains (Heil *et al.*, 1992). It has also been shown that some neurons respond only to a particular call type common to the species being studied. This means that higher level processing of the temporal changes in spectra is occurring in the avian brain.

The following material describes bird audition from a behavioral point of view as discussed by Dooling (1982). Although birds generally have higher hearing thresholds than humans, certain owls have much more acute hearing. As with humans there is good correlation between the frequencies used in vocalization and those at which hearing is most acute in songbirds (1.5-2.5kHz). Non-songbirds (e.g. owls) do show differences which may be related to the optimization of hearing for tasks such as hunting.

Frequency patterns are used by birds in song recognition. It has been shown that frequency (f) discrimination thresholds in birds (measured as $\Delta f/f$) are lowest from about 1-4kHz. Humans are two to three times as sensitive to frequency changes as birds, and are so over a wider frequency range. A bird's ability to discriminate intensity or temporal differences does not differ much from that of mammals according to Dooling (1982).

It is difficult to determine the exact cause and effect relationships between vocal/hearing abilities of animals and the purposes for which they use sounds. Even so, it may be that birds compensate for poorer hearing sensitivity and selectivity by producing louder and more frequency modulated sounds than humans do with speech. Human communication contains many subtleties, perhaps because a fairly large amount of information is being coded; there may have been selective pressure in humans for more sophistocated hearing, in order to achieve better information transfer. It may also be true that birds use song for long-range communication under conditions where visual cues are obscured (S.E. Cosens, pers. comm.).

1.1.4.3 Synthesis

Vocalizations in birds are produced in a similar manner to human speech. Studies of bird audition at neuro-physiological and behavioral (equivalent to psychological) levels indicate that hearing in birds is very similar to that of humans. If anything, birds appear to

be less adept at hearing (excluding owls) than we are. In fact, one might surmise that both the intensity and the frequency sweeps in bird calls may be exaggerated to compensate for inadequacies of the avian ear. Both birds and humans appear to sense the frequency on a ratiometric scale. This indicates that the spectral transformation partitions frequency logarithmically.

1.1.5 Neural processing of sound

Our understanding of biological bases for temporal integration in animals is sketchy. In vertebrates, we do have detailed neurological information on some aspects of human speech, on hearing in Barn Owls and on echolocation in bats.

In the Mustached Bat, for example, spectral resolution is maximal around the echo frequency and minimal at the emitted pulse frequency (Doppler shift results in the frequency difference). We know that a Barn Owl processes time-of-arrival information in parallel with intensity information to localize its prey (Konish, 1993). It later integrates time, intensity and frequency information to form an auditory map of prey location. Similarly, bats process Doppler shift information, time-of-flight for pulses, and echo intensity information (Suga, 1990). With it they are able to identify prey location, air speed, target size and approximate identity. The bat is also able to vary its pulse frequency, through feedback, to maintain the echo within its most sensitive hearing range.

Although the processing algorithms that bats and the owl use are well understood, the neural mechanisms are not. One area which has shown much promise is the discovery of delay lines and coincidence detectors used by both animals to map time delays. In the bat (Suga, 1990), specific neurons perform a logical AND operation on the inputs from two neurons. A neural pulse due to an echo is fed to AND neurons via a binary tree to equalize delays, while the emitted neural pulse is routed to the same AND neurons through various lengths and numbers of neurons. This introduces a distribution of delays. Each AND neuron has its best delay (BD). These neurons are called delay-tuned neurons (they are in fact ensembles of neurons). Apparently the animal learns to map activity within the array of AND neurons to distance.

It seems that animals have developed, through evolution, specialized feature extractors which provide general information about timing, intensity and spectral composition of signals. These feature extractors operate in parallel. Their outputs are combined to extract more abstract features which are the symbols of communication, location of prey, and related parameters.

Suga (1990) has stated that: "It seems likely that the human auditory system also uses such neural arrays to process speech sounds. Even though the processing of speech is unique and predominantly a function of the neocortex, the subcortical auditory regions, with their combination-sensitive neurons, may analyze the sounds of speech to a greater extent than has been generally thought". If this is so, an artificial system based on similar principles, could be also be built. Dror et al. (1995) for example developed a sonar target recognition device using an artificial neural network and a bat echolocation model.

In the Mustached Bat, objects that are close to the bat reflect a sonar pulse with little delay, while those that are distant are represented by longer delays. Delays of 2 to 28 msec represent distances of 30 to 500 cm. Until recently, it was not known how this ensemble of delays received over 30 msec was assembled into an image of the bat's environment. Dear et al. (1993) discovered a neural mechanism in the brown bat for image assembly. They found neurons which had maximum sensitivities at different BD's, but also found neurons which fired long after they were stimulated by a delay close to their BD. This latency was inversely proportional to the neuron's BD. Since neurons sensitive to short delays had long latencies, images of progressively more distant scenes would be formed one after the other, until the final one (at about 30msec) was a picture of the whole scene within the bat's auditory field.

Further evidence was consistent with this neural process being a mechanism for scene assembly. Neurons develop narrower response curves (higher Q) over time. This means that neurons with short BD's have the best delay resolution. Thus, as expected, the resolution of the image for near objects (short delays) should be better than that of distant ones. A similar technique, multi-resolution decomposition, has been used in engineering for image analysis and object recognition among other applications.

1.2 Pattern recognition

The task in pattern recognition is to find a means to automatically recognize or classify something based on data from or about it. There are several challenges in this process. One must first acquire the raw data. Next, raw data must be processed into a more useful form for analysis (pre-processing). Lastly, some analysis technique must be used to generate the desired output. In the case of identifying birds by their sounds, data can be acquired easily with a directional microphone and sound recording equipment. The task of specifying appropriate pre-processing and analysis methods is not simple.

In order to properly address the pre-processing issue, a section has been devoted to the explanation of several commonly used methods. Similarly, data analysis naturally involves traditional statistics and the newer field of artificial neural networks (ANNs). An overview of pertinent methods is presented for each type of analysis. In addition, characteristics of statistical methods and ANNs are compared.

1.2.1 Pre-processing

Building on knowledge that vertebrates, and more specifically birds, produce and respond to sounds that have complex temporal and spectral characteristics, it is necessary to examine pre-processing techniques that can be used to quantify features. The purpose of this section is to provide a theoretical background on some signal processing techniques, with discussions of biologically relevant features being left for later chapters.

Measuring temporal parameters of bird sounds is relatively simple. Obtaining frequency information, however, requires the use of more complex techniques. Although there are many transformations that can convert a time-domain signal into a frequency-time

representation, including the more common Fourier and wavelet transforms (Hlawatsch and Boudreaux-Bartels, 1992), Fourier analysis was chosen as a first step in this research, due to its wide use and conceptually simplicity. A natural extension of Fourier analysis involves obtaining statistically sound estimates of spectral content by calculating power spectral densities (PSD).

Other pre-processing methods are inspired by the source filter model commonly applied in speech analysis. The last two parts of this section are thus devoted to developing linear predictive coding analysis upon a foundation of relevant digital filter theory.

1.2.1.1 Discrete Fourier transform

The discrete Fourier transform (DFT) may be expressed in one of two ways. Both provide a fully reversible transform (Oppenheim and Shafer, 1975):

$$F_k = \sum_{j=0}^{N-1} e^{ijk2\pi/N} \cdot f_j \Leftrightarrow f_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{-ijk2\pi/N} \cdot F_j \,, \qquad [1]$$

or (Papoulis, 1980):

$$F_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{-ijk2\pi/N} \cdot f_j \Leftrightarrow f_k = \sum_{j=0}^{N-1} e^{ijk2\pi/N} \cdot F_j \,, \qquad [2]$$

where $i = \sqrt{-1}$ and j,k are integer indices.

The first form will be used in this document. A DFT of purely real data results in complex Fourier coefficients with the property (Press et al., 1992):

$$F_{N-n} = (F_n)^* \,, \qquad [3]$$

where the asterisk indicates the complex conjugate operation.

Thus for real data, we can obtain all the relevant information about the signal with the N/2+1 values $F_0, F_1,...,F_{N/2}$. The remaining N/2 can be thought of as 'negative frequency' coefficients, and are of the same magnitude as the corresponding positive frequency coefficients. $F_{N/2} = F_{-N/2}$ and F_0 are always real.

Parseval's Relation for the DFT states that the total power contained in a finite time domain sequence is the same as that contained in the Fourier transform of that sequence. Using eq. [1] for the DFT, Parseval's Relation is of the form indicated below. The mean squared amplitude in the time domain equals the sum of the squared harmonic magnitudes, where the magnitudes are divided by N (after Press et al., 1992; Oppenheim and Shafer, 1975):

$$\sum_{j=0}^{N-1}|f_j|^2 = \frac{1}{N}\sum_{n=0}^{N-1}|F_n|^2 \quad \text{or}$$

$$\frac{1}{N}\sum_{j=0}^{N-1}|f_j|^2 = \frac{1}{N^2}\sum_{n=0}^{N-1}|F_n|^2 = \sum_{n=0}^{N-1}\left[\frac{|F_n|}{N}\right]^2. \qquad [4]$$

When one applies the DFT, it is invariably in the form of a fast Fourier transform (FFT). FFT algorithms have been designed for computational efficiency (Press et al., 1992). One problem with the DFT is that there is leakage of spectral content between coefficients. Explicit windowing of time domain data is usually performed to reduce leakage.

1.2.1.2 Time window functions

Spectral leakage occurs because a finite sequence generally has an abrupt beginning and end. Effectively, an infinite sequence in time is being multiplied by a rectangular window of finite width before the Fourier transform is applied. In the frequency domain, this amounts to convolving a sinc-like function with the transform of the data. The rectangular window function and its Fourier transform are (Oppenheim and Shafer, 1975):

$$w_R(m) = \begin{cases} 1, & |m| < N \\ 0, & otherwise \end{cases} \Leftrightarrow W_R(\omega) = \frac{1}{N}\left(\frac{\sin[\omega(2N-1)/2]}{\sin[\omega/2]}\right). \qquad [5]$$

The side-lobes of this rectangular window spectrum have considerable magnitude. When the original data are multiplied by such a function, narrow spectral peaks are widened. It is better to use explicit windowing, where the window function has been designed to have smaller side-lobes. A rather simple window with this characteristic is the triangular Bartlett window (Oppenheim and Shafer, 1975):

$$w_B(m) = \begin{cases} \dfrac{N-|m|}{N}, & |m| < N \\ 0, & otherwise \end{cases} \Leftrightarrow W_B(\omega) = \frac{1}{N}\left(\frac{\sin[\omega N/2]}{\sin[\omega/2]}\right)^2 \qquad [6]$$

Time-domain coefficients for the Bartlett window (w_j) are also simple to compute (Press et al., 1992):

$$w_j = 1 - \left|\frac{j-0.5N}{0.5N}\right|. \qquad [7]$$

1.2.1.3 Power Spectral Density

The power density spectrum indicates the distribution of power in a signal with respect to frequency. In discrete systems, one can both smooth spectra and reduce the number of frequency bins using PSD estimates. One way to produce a PSD is to use a periodogram. The periodogram is a single-sided power spectrum that is in keeping with Parseval's relation (Press et al., 1992):

$$P(F_0) = \frac{1}{N^2}|F_0|^2$$

$$P(F_k) = \frac{1}{N^2}\left[|F_k|^2 + |F_{N-k}|^2\right] \qquad k = 1,2,\ldots,\frac{N}{2}-1 \qquad [8]$$

$$P(F_{N/2}) = \frac{1}{N^2}|F_{N/2}|^2$$

The mean squared amplitude thus equals the sum of $P(F_k)$, for k = 0...N/2.

A problem with this method is that $P(F_k)$ are not good estimators of the true power values. The reason is that the standard deviation associated with the mean of $P(F_k)$ is equal to $P(F_k)$. This cannot be remedied by sampling more points. The degrees of freedom that one might have expected to gain by increasing the number of time domain samples M, only increases the number of frequencies for which one has a power estimate. There are two ways to rectify this problem.

The first technique is to calculate the $P(F_k)$, using a DFT of size N = M. The $P(F_k)$ are then grouped into a number of bins L (where L<M), and summed within the bins (summing rather than averaging preserves Parseval's relation). The new $P(F_j)$, with j = 0...L, represent the power at the midpoints of their respective bins. $P(F_j)$ have a variance which is now reduced by a factor of M/L.

In the second technique (Bartlett's procedure), the M samples are divided into K segments and $P(F_k)$(using DFT's of size N = M/K) of each segment are calculated. The K powers for a given frequency ($P(F_k)$) are averaged to yield $\overline{P}(F_k)$. Again, the variance associated with $\overline{P}(F_k)$ is reduced by a factor of K compared with the variance of $P(F_k)$. This technique allows smaller amounts of sampled data to be handled at one time. With the FFT, computational effort is of order Nlog(N). Computing several small FFTs is more efficient than computing the single large one of the first method. It is thus better to use Bartlett's procedure. Note that true power representation would require using the sum, rather than the average of K values; the description here is in keeping with Oppenheim and Shafer (1975).

The Welch method of estimating the PSD is simply Bartlett's procedure preceded by explicit time domain windowing of the data segments. In order to maintain Parseval's Relation with windowing, one may generalize the PSD by changing the normalization factor (Press et al., 1992):

$$P(F_0) = \frac{1}{W_{ss}^{\,2}}|D_0|^2$$

$$P(F_k) = \frac{1}{W_{ss}^{\,2}}\left[|D_k|^2 + |D_{N-k}|^2\right] \qquad k = 1,2,\ldots,\frac{N}{2}-1$$

$$P(F_{N/2}) = \frac{1}{W_{ss}^{\,2}}|D_{N/2}|^2 \qquad [9]$$

where

$$D_k = \sum_{j=0}^{N-1} e^{ijk2\pi/N} \cdot w_i \cdot f_j,$$

$$W_{ss} = N\sum_{j=0}^{N-1} w_j$$

Note that when the window function is rectangular, W_{ss} becomes N^2, and we have the original set of equations for the PSD. To compute the PSD using the Welch method, one simply segments the time domain samples, windows them, performs the K FFTs and averages like coefficients.

The way that various literature sources define and display PSD and spectra is inconsistent. Although care was taken to be consistent in the formulae above, the goal in this thesis was to find an accurate estimate of spectral content with a small number of bands. The estimates should form a smoothed spectral representation with reduced dimensionality. For this purpose, the Welch method can be used to find the average power within N bands whether the input sequence is long or short. It should be emphasized that these are band estimates. For real value inputs, redundant negative frequency values need not be used. The 'zero frequency' value actually represents frequencies from DC to ± fs/2N, where fs is the sampling frequency, and thus should be retained. Taking the square root of the N/2+1 positive and 'zero frequency' estimates yields properly averaged magnitudes $S(F_k)$ for analysis (Press et al., 1992):

$$S(F_k) = \sqrt{\frac{1}{KW_{ss}^{\,2}}\sum_{i=1}^{K}|D_{ki}|^2} \qquad k = 0,1,\ldots,\frac{N}{2}$$

where D_k *and* W_{ss} *are as above and* [10]

$$w_j = 1 - \left|\frac{j-0.5N}{0.5N}\right|$$

That $S(F_k)$ provide good spectral estimates can be verified by testing the technique with a random sequence (a uniform distribution with range ±0.5) as input. As expected, average magnitudes $S(F_k)$ are about equal, and variance decreases as the random sequence of length M becomes longer for constant N (Fig. 1.5).

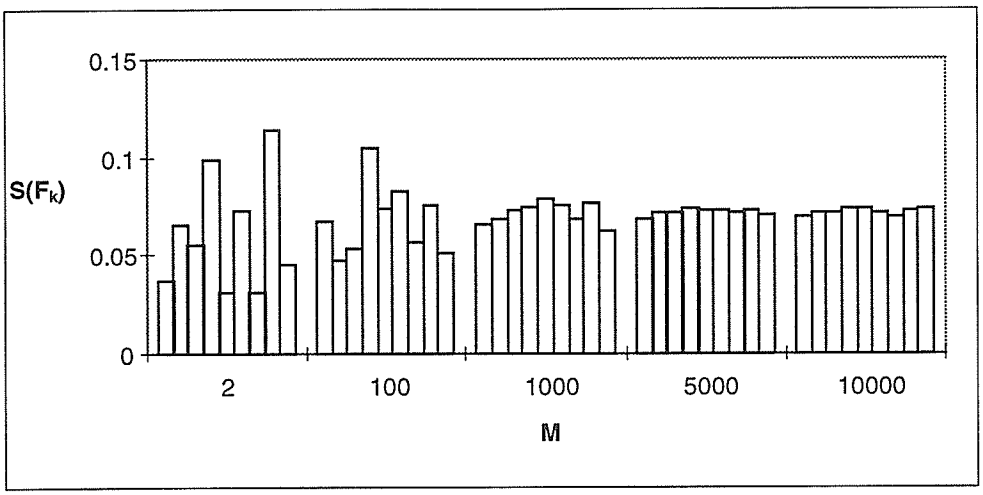


Fig. 1.5 Effect of increasing M on the PSD of a random signal, given a constant FFT size (N=16). Bars represent the average magnitude within the nine non-negative frequency bands.

1.2.1.4 Digital filters

This section is intended to provide a brief overview of digital filters in order to support the section on linear predictive coding analysis; information for the overview was derived from several references (Oppenheim and Shafer, 1975; Papoulis, 1980; Oppenheim et al., 1983). Since all data used in this study are sampled, discussions will be restricted to theory involving the z-transform, which is the discrete time analog of the Laplace transform. The unilateral z-transform can be specified as:

$$X(z) = \sum_{n=0}^{\infty} x[n] z^{-n}, \qquad [11]$$

where x(n) is a sequence of numbers, and:

$$z = re^{j\omega}, \qquad [12]$$

is a complex exponential variable. If the radius is equal to one, the transform becomes the Fourier transform of x(n):

$$X(z) = \sum_{n=0}^{\infty} x[n] z^{-n} = \sum_{n=0}^{\infty} x[n] r^{-n} e^{-j\omega n}. \qquad [13]$$

When only a finite number of points (N) are non-zero, one can also specify the z-transform as a finite sum. In this section, it is the application of finite z-transform to filter

theory which is of interest. The convolution property of this transform makes the study of digital filters simpler:

$$x_1[n]* x_2[n] \xleftrightarrow{Z} X_1(z) \cdot X_2(z), \qquad [14]$$

where the convolution sum is defined as:

$$x_1[n]* x_2[n] = \sum_{k=0}^{N-1} x_1[n] \cdot x_2[n-k] \ . \qquad [15]$$

The transfer function of any filter is determined by the relationship between its input and output in the frequency domain; if the input spectrum is X(z) and the output is Y(z), the transfer function is H(z)= Y(z)/X(z). It is then apparent that a filter can be described in the frequency domain as:

$$Y(z) = H(z) \cdot X(z). \qquad [16]$$

If one knows the spectrum of the transfer function, it is relatively simple to implement a filter in the time domain using the convolution property of the *z*-transform:

$$y[n] = h[n]* x[n] = \sum_{k=0}^{N-1} h[n] \cdot x[n-k],$$
$$H(z) = \sum_{k=0}^{N-1} h[n] z^{-n}. \qquad [17]$$

For general case where N = ∞ these equations describe the most basic digital filter, the finite impulse response (FIR) filter. The magnitude of the coefficients decays exponentially as one moves out from the center tap(s) of the filter. Note that practical implementations with a finite number of coefficients are an approximation of the general case. Since the output depends only on the present and past inputs x[n-k], one can apply an impulse to the input of a practical filter (N << ∞) and have the effect of the spike last for finite time at the output. This assumes that coefficients have been chosen such that the filter is stable. Even with the general case, exponential decay among coefficients in a stable filter implementation ensures that the effect of the impulse declines with time. Due to the relationship between the finite *z*-transform and the DFT, performing an FFT on the FIR filter coefficients provides a simple means of examining spectral characteristics if the filter's impulse response.

FIR filter implementations are achieved by convolving the filter coefficients h[n] with the samples. This may also be thought of as a multiplying samples from n time units in the past by the appropriate coefficient, hence z^{-0},z^{-1}, z^{-2},...,z^{-n} represent time delays of 0,1,2,...,n. The FIR filter can also be expressed in polynomial form by changing the notation for h[n] to b_i:

$$Y(z) = (b_0 \cdot 1 + b_1 \cdot z^{-1} + ... + b_n \cdot z^{-n}) X(z), \qquad [18]$$

which leads to the polynomial form of the transfer function:

$$H(z) = (b_0 \cdot 1 + b_1 \cdot z^{-1} + ... + b_n \cdot z^{-n}). \qquad [19]$$

Note that the roots of this polynomial are all points at which the equation is zero. This is thus called an all-zero filter. The more general IIR filter, contains both zeros and poles (zeros in the denominator):

$$H(z) = \frac{(b_0 \cdot 1 + b_1 \cdot z^{-1} + ... + b_n \cdot z^{-n})}{(1 + a_1 \cdot z^{-1} + ... + a_n \cdot z^{-n})}. \qquad [20]$$

Its response contrasts with that of FIR since the output depends on both past inputs and outputs. This means that an input spike affects the output forever.

If the numerator of the IIR filter is unity, one is left with an all-pole filter. This is the type of filter extracted in the process of linear predictive coding (LPC). Note that the more general IIR filter becomes a FIR filter if the denominator is unity.

1.2.1.5 LPC

The source-filter model of speech production provides a good theoretical basis for analysis of speech and speech-like signals; theory described here is derived from several sources (O'Shaughnessy, 1989; Atal, 1985; Ingle and Proakis, 1991). An all-pole filter model can be used to represent a vocal tract that contains concatenated tubes, but no side branches. The vocal tract is believed to be relatively constant in shape within 10-30 msec windows (O'Shaughnessy, 1989). For this reason, sampled speech waveforms are usually analyzed with windows of this length. Linear predictive coding analysis (LPC) is frequently used to extract all pole filter coefficients from speech windows. The LPC transfer function is:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{N} a_k \cdot z^{-k}}, \qquad [21]$$

where there are N coefficients a_k (usually 10-12) and G is the gain of the filter. Note that the sign of the sum in some texts is negative rather than positive as was described for the IIR filter above. This seems to be a notational convention, and does not affect spectral magnitudes of the filter function. I follow Ingle and Proakis (1991) in making the sum positive.

If one lets G=1, and defines P(z)=1/H(z), one notices that the effect of the filter H(z) can be canceled by filtering with P(z): X(z)=P(z)H(z)X(z). H(z) is the filter that models the vocal tract, while P(z) is a 'whitening' filter which can be used to cancel the effect of the vocal tract and leave the excitation as a residual. It also turns out that the whitening filter is an FIR filter:

$$P(z) = \frac{1}{H(z)} = 1 + \sum_{k=1}^{N} a_k \cdot z^{-k}. \qquad [22]$$

This means that one can obtain the inverse of the speech envelope by first finding the all pole coefficients, performing a DFT on them, and taking the reciprocal of the resulting magnitudes. If the result of this process is superimposed on the DFT results for the raw samples, one sees that LPC provides a smoothed representation of the spectral envelope (Fig. 1.6).

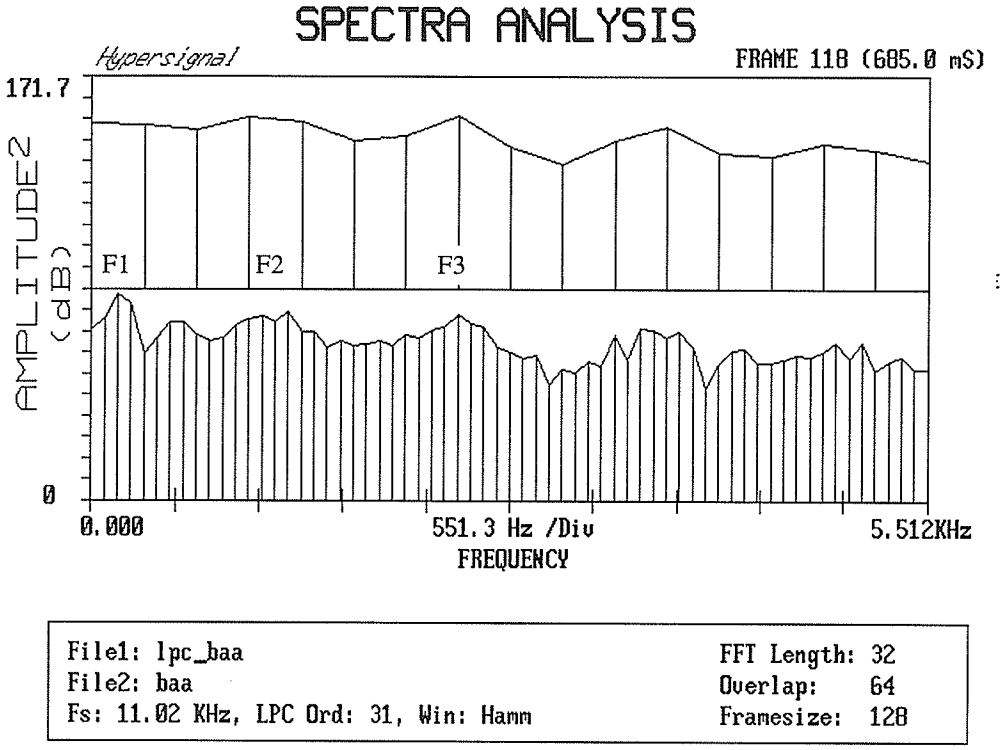


Fig. 1.6 LPC analysis of a window from the vowel sound in 'baaa'; top is FFT of the 32 LPC coefficients, bottom is the raw data. The same three formants are indicated as they were in the spectrogram.

Speech coding is often done by first extracting coefficients of H(z) for windows of speech. Then one uses the whitening filter P(z) to remove the speech envelope. What remains in the residual after 'whitening' is an approximation of the speech excitation. From this residual, one can then determine whether or not the excitation is aperiodic (unvoiced) or periodic (voiced). If it is voiced, the pitch of voiced excitation can be determined as well. Extracting the pitch and filter information in this way permits the speech to be transmitted or stored in a condensed form. Speech can be synthesized by generating white noise, or pulses with the appropriate pitch and then filtering the excitation with the LPC filter H(z).

In reality, this technique has drawbacks. For example, nasal consonants produced by the introduction of a branched tube into the vocal tract introduce zeros that are poorly

captured by the filter. This may necessitate the use of more coefficients to approximate zeros using only poles. The technique is also computationally intensive. The calculation of coefficients is performed using either autocorrelation analysis of the samples, or by calculating filter reflection coefficients, and then recursively calculating the coefficients. These problems aside, LPC provides a powerful tool for analyzing signals that are easily characterized by a relatively unbranched vocal tract model; a worked example of envelope determination is included in the Appendices.

1.2.2 Statistical methods

This section provides a rationale for the choice of multivariate methods used in the thesis. Most of the information was derived from SAS (1985) and Manley (1994). Before proceeding, two definitions are required for the sake of clarity. First, the term 'hypothesis testing' refers to the process of determining whether or not there is sufficient evidence to refute a proposed hypothesis. The term 'prediction' refers to the ability of a method to interpolate or extrapolate from a small sample of records to the larger population. Fig. 1.7 provides a general framework for choosing the data analysis methods describe in this section. Artificial neural network (ANN) techniques are discussed in detail in a separate section.

1.2.2.1 Regression

Regression can be used for both hypothesis testing and prediction. The simplest form is linear regression, which can be thought of as the process of estimating parameters β_i in the following model (Neter and Wasserman, 1974):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad [23]$$

where Y is the dependent variable (response), X the independent variable and ε a random error term for the *i*th observation. The mean (expected value) and variance of the error are $E(\varepsilon_i) = 0$ and $\sigma^2(\varepsilon_i) = \sigma^2$ respectively. The error covariance $\sigma(\varepsilon_i,\varepsilon_j)$ is zero, meaning that the errors are uncorrelated. β_0 is often called the intercept and β_1 the slope of the regression line.

This model draws attention to a number of features of the model which also apply to the other methods in the General Linear Model family including ANOVA (see below). First, the parameters and error terms additively determine the response. With the assumption of zero error mean, the regression function thus becomes:

$$E(Y) = \beta_0 + \beta_1 X \qquad [24]$$

which is a mapping of the probability distribution of X to that of Y.

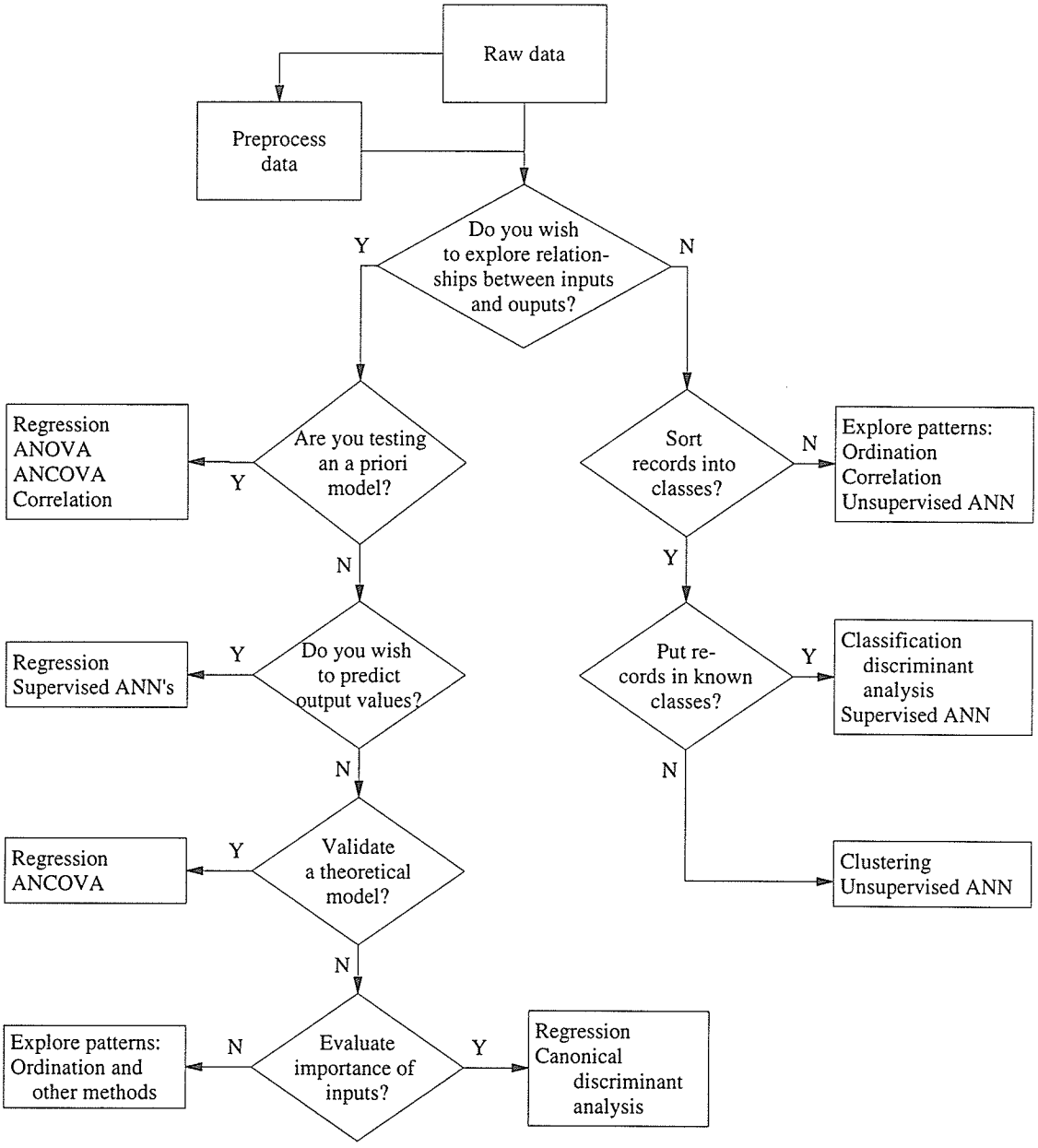


Fig. 1.7 A process for choosing relevant analysis methods for a given data set and research goal.

Another important assumption is that error variance is relatively constant over the range of X. In order to use linear regression with confidence, this assumption should be checked by examining the residuals (ε_i) over the range of X. If one adds the additional assumption that ε_i are normally distributed, it becomes possible to test hypotheses and establish confidence limits for both parameters and predicted values of Y. Although the assumption of normality is weaker than that of homogeneity of variance, one should verify that the residuals are at least somewhat normal in their distribution (Neter and Wasserman, 1974).

The parameters of this model are determined using the method of least squares. The overall variation in Y, represented by the total sum of squares (SSTO), can be partitioned into the component of total variation from the regression model (SSR) and the remainder SSE. The idea is to minimize the sum of squared errors (SSE),

$$SSE = \sum_i \{Y_i - \beta_0 - \beta_1 X_i\}^2 \qquad [25]$$

over all observations. The term within the sum is the predicted value of Y subtracted from the actual value of Y. By differentiating SSE w.r.t. β_0 and β_1 , setting both derivatives to zero, and substituting particular parameter values b_0 and b_1 , one is left with a system of two simultaneous equations to solve:

$$\begin{aligned} &\sum_i Y_i - nb_0 - b_1 \sum_i X_i = 0, \\ &\sum_i X_i Y_i - b_0 \sum_i X_i - b_1 \sum_i X_i^2 = 0, \end{aligned} \qquad [26]$$

The unique solution for b_0 and b_1 is:

$$\begin{aligned} b_1 &= \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}, \\ b_0 &= \overline{Y} - b_1 \overline{X}. \end{aligned} \qquad [27]$$

For details of this derivation, see Neter and Wasserman (1974).

In regression analysis, one usually proposes a model such as the one above and estimates the parameters. One can then analyze residuals to verify that the model's assumptions are being met. If they have been met, one can assess the statistical significance of the model, and make interpolations or extrapolations using it. If assumptions have not been met, other options are available. It may be helpful to modify the model, transform one or more variables (i.e. with log or square root functions), or remove seemingly erroneous data (outliers). Several independent variables can also be used (multiple linear regression), including non-linear transforms (e.g. polynomial) of the original variable(s).

Linear regression requires that the model is "linear in its parameters". For example, linear regression can be applied to data that contains an exponential relationship between variables by specifying the model using logarithms. Even with such flexibility, there are still times when linear regression is inappropriate due to strong non-linearities in the data. In such cases, other techniques like iterative nonlinear regression or autoregression (for time series data) are available.

1.2.2.2 Analysis of variance (ANOVA)

ANOVA is much like linear regression, with the difference that independent variables must be discrete. A wide variety of ANOVA designs are available for the purpose of

hypothesis testing. The simplest involves testing the hypothesis that the mean value of a variable is the same for two groups; states of the independent variable are often called treatment levels. This could be accomplished using a simple Student's *t*-test or by proposing a linear model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad [28]$$

for which X is zero for one treatment level and one for the other. It is possible to have complex designs that have several treatments, each with several treatment levels. If several variables are believed to contribute to differences among treatments, multivariate analysis of variance (MANOVA) may be appropriate.

When there are more than two levels of a treatment and the means for some levels differ, ANOVA only tells us whether or not one of the means is different from at least one other. It may be interesting to determine where differences lie. For this purpose, multiple comparison techniques have been devised that allow one to determine the overlap of responses to treatment(s). Some methods include those of Tukey, Bonferroni and Scheffé; the Scheffé method is most useful with unbalanced designs (those with unequal sample sizes for levels) where one does not wish to explore a limited number of specific comparisons (Neter and Wasserman, 1974).

Analysis of covariance (ANCOVA) is a hybrid between ANOVA and regression. In fact, ANOVA, regression and ANCOVA are just special cases of a General Linear Model. Complex designs and models are usually handled with software capable of implementing a General Linear Model. It is important to assess the validity of assumptions with all of these methods.

1.2.2.3 Classification discrimination

These discriminant methods are used to classify unknown observations from a population based on known observations drawn from the same population. If the within-class probability distributions are approximately normal, classificatory discriminant analysis is appropriate. The usual method of performing this analysis is often termed 'linear discriminant analysis', when a common covariance matrix is used for all classes. When covariance matrices for classes differ substantially, 'quadratic discriminant analysis' is used. Software like that of SAS (1985) tests for homogeneity of covariance and chooses the appropriate method of analysis.

To understand this method, begin by assuming that there are *t* groups or classes of observations. Observations within a group are represented by a vector of variables **x**, with group mean vector $\mathbf{m}_t$ and covariance matrix $\mathbf{S}_t$. It is assumed that the means and covariance matrices are parameters of a multivariate normal distribution. Given these assumptions, the distance between an observation vector and each of the mean vectors can be measured using the squared Mahalanobis distance:

$$g_1(x,t) = (x - m_t)^T S^{-1} (x - m_t). \qquad [29]$$

This is a simple linear discriminant function, which contains the assumptions that all S_t are identical, and that prior probabilities (priors) for the occurrence of each class are equal. The classification rule assigns an observation x_i to the class for which the squared distance is smallest (Manley, 1994).

SAS (1985) allows more flexibility by using the generalized squared distance:

$$D_t^2(x) = g_1(x,t) + g_2(t). \qquad [30]$$

Note that this reduces to the Mahalanobis distance when all covariance matrices are the same, and prior probabilities (q_t) are equal ($g_2 = 0$). If prior probabilities are unequal, enhanced linear discrimination is achieved by setting:

$$g_2(t) = -2\ln(q_t). \qquad [31]$$

In quadratic discrimination, g_1 becomes:

$$g_1(x,t) = (x - m_t)^T S_t^{-1}(x - m_t) + \ln|S_t|. \qquad [32]$$

The classification rule, when generalized squared distances are used, is to classify a given observation into the class u for which it has smallest distance $t = u$ or the highest posterior probability of membership. This probability is:

$$p_t(x) = \frac{\exp(-0.5D_t^2(x))}{\sum_u \exp(-0.5D_u^2(x))}. \qquad [33]$$

Readers familiar with pattern classification will recognize references to Bayes decision theory (Duda and Hart, 1973). In effect, this form of discriminant analysis is performing Bayesian classification for a special case where multivariate normality is a reasonable assumption. Given an a priori probability of a class being present in a sample, and a conditional probability of the sample belonging to a given class, the posterior probability of that class being the correct choice can be determined:

$$P(t|x) = \frac{p(x|t)P(t)}{p(x)} \qquad [34]$$

P(tlx) is the probability that the observation belongs to class 't' given the sample data (posterior probability). p(xlt) is the probability density function for samples 'x' belonging to a given class, and in linear discriminant analysis are determined by the means and covariances. P(t) is the a priori probability of observing a member of class t and p(x) is a normalizing factor that guarantees that posterior probabilities sum to one:

$$p(x) = \sum_n [p(x|t)P(t)]. \qquad [35]$$

The Bayesian decision rule is then to classify the observation into the class that results in the largest posterior probability. For an extensive examination of Bayesian decision theory and related pattern classification techniques see Duda and Hart (1973).

1.2.2.4 Stepwise procedures

Algorithms are available that help the users to select a subset of variables from their input data set to use in subsequent analyses. These are applied to regression and classification discrimination tasks (SAS, 1985). These techniques add variables to or remove them from a predictive model in steps. Criteria for adding or deleting a variable may be based on significance levels from F-tests for an ANOVA or ANCOVA model or partial correlation coefficients. Three methods are common.

In forward selection, a variable is added at a given step if it appears that it will make the largest improvement (SAS, 1985). In backward elimination, all variables are initially entered into the model, and those that appear to make the least contribution are progressively removed. Some stepwise procedures successively add variables, and test their contribution. If their contribution becomes insignificant, they are later removed.

There are several reasons for applying care in the use of such procedures. The user must choose the point at which he or she feels that the optimum number of variables has been selected. The set of variables chosen is not always the 'best' set, especially if the algorithm is inflexible (e.g. simple backward or forward methods). Methods that permit variables to be entered and removed are probably better in this respect. It is also reasonable to assume that the results will depend on the entry criteria and choice of significance levels used to control entry and removal of variables. James and McCulloch (1990) emphasize that good predictive variables selected by stepwise methods may not be of any theoretical importance; great care must be taken in making statements about cause-effect relationships. Stepwise procedures can be valuable for variable selection in pattern recognition applications, if they are used carefully.

1.2.2.5 Ordination

The term 'ordination' is used commonly in biological literature, and refers to the process of finding a representation for multidimensional data (Pielou, 1984) in fewer dimensions than the data space occupies. There are many methods available for this task, each with its own assumptions, advantages and disadvantages. Principal Component Analysis (PCA) is a computationally and intuitively simple method for ordination. The following explanation of PCA is derived from several references (Mardia et al., 1979; Pielou, 1984; Srivastava and Carter, 1983).

The main applications of PCA are in dimensionality reduction, data compression and data exploration. PCA is helpful when data for several variables are measured in a study, and one wishes to view trends in the data in a lower dimensional space (usually one or two dimensions). Mardia et al.(1979) illustrate the process by analyzing a set of five equally weighted exam scores with the goal of comparing student performance. In this type of

situation, one usually obtains an overall score by calculating the mean of the exam mark for each student. This can be thought of as multiplying each mark by a weight and summing the results:

$$o = w' \cdot i \qquad [36]$$

where w' = {0.2, 0.2, 0.2, 0.2, 0.2}, **i** is a column vector of inputs and o is the output score. If the sum of squares of the weights is 1.0 then the score is called a standardized linear combination (SLC) of the marks. The variance about the mean score is not necessarily maximal; one could argue that the best comparison of student performance is to spread the final scores as much as possible (i.e. to find the linear combination of marks that gives the largest variation in score). From a dimensionality reduction point of view, this means that we want to find the one score that summarizes as much of the variation in the unreduced data set as possible.

The PCA process can be thought of as an exercise in finding first the SLC which has maximal variance, then the next SLC which accounts for the largest remaining variance until all variance is accounted for. These scores are uncorrelated.

In practice, PCA axes are extracted using eigenanalysis. One begins by considering the Jordan/spectral decomposition theorem which states that any symmetric matrix S can be written as:

$$S = PVP^T, \qquad [37]$$

where P is an orthogonal matrix containing columns which are standardized eigenvectors of S; V is diagonal and contains the eigenvalues of S. The theorem can be rearranged to obtain:

$$P^TSP = V. \qquad [38]$$

Given the matrix of marks X, with mean vector **u**, and covariance matrix S, we seek a transformation of X to output Y such that the output scores are uncorrelated. Given that the columns of X are the marks for the five exams, columns of Y are called the principal components of X:

$$Y = (x - \boldsymbol{u})P^T, \qquad [39]$$

with V containing the eigenvalues of S in descending magnitude from left to right. Because we have transformed S into a diagonal matrix V, all covariances become zero and the eigenvalues represent the variance in their respective scores. The lack of covariance terms emphasizes the fact that the PCs are uncorrelated. Total variance is equal to the trace of V. An example calculation of PCA is included in the appendices.

It is worth noting that a correlation matrix can be substituted for the covariance matrix (Manly, 1994). The method is still called PCA, but the eigenvalues now represent standardized variances. This change automatically standardizes input variables so that

each has equal weight in the analysis. This is particularly useful when variables are measured on vastly different scales.

Some other methods related to PCA include factor analysis, principal coordinates analysis (PCOA), canonical correlation analysis, correspondence analysis and canonical discriminant analysis. The later method is of particular interest and is described below. For data that has strongly non-linear trends, even with transformation, non-metric multidimensional scaling (NMDS) can be a powerful tool; although NMDS is non-parametric (distribution free), it is an iterative method. It thus may yield sub-optimal local solutions, and may require extensive computational resources.

1.2.2.6 Canonical discrimination

Canonical discrimination is useful for dimensionality reduction and data visualization. It is related to PCA, but allows the identity of data labels to provide additional information (supervision). One can interpret the result as being the production of variables which best separate the classes (Manley, 1994). Assuming that all classes share a common multivariate normal distribution with a common covariance matrix, analysis proceeds as follows (SAS, 1985):

1. Transform raw input variables to make the covariance matrix (pooled within-class) the identity matrix.

2. Calculate the mean vectors for each class.

3. Calculate weighted means based on the number of samples in each class, and perform PCA on the means. The magnitude of the eigenvalues now represents the ratio of the interclass variation to the intraclass variation. The direction of the variation, is given by the PCA axes.

4. Transform the PCA scores back to the original data space. The resulting variables are now the canonical discriminant variables. Note that these variables are uncorrelated, but not perpendicular.

Manley(1994) suggests that this method may still be valuable, even when the assumptions of normality and equal class covariances are violated. In these cases, the optional application of the method for hypothesis testing is pre-empted.

1.2.2.7 Clustering

Unlike discrimination, clustering is mainly used to find groups of similar records where the groups are unknown prior to analysis. Groups are called clusters. Manley (1994) subdivides clustering into hierarchical and partitioning methods. Hierarchical methods easily provide a dendrogram, which is a tree-like graphical representation of groupings. Divisive hierarchical clustering algorithms begin with a cluster that contains all the records, and progressively divides it into groups of smaller and smaller clusters. Agglomerative algorithms begin with individual records and progressively group them

together. In partitioning methods, one chooses the number of groups before analysis, and algorithms iteratively add and remove records from the groups. Groups are also moved in the data space to improve the representation. Clustering methods can use many measures of similarity (or difference) to group records. Another name for cluster analysis is unsupervised pattern recognition (Everitt, 1993).

One can devise a simple cluster analysis algorithm using Euclidean distance (Manley, 1994). The first task is to compute a distance matrix for all record pairs (Table 1.1). Since the matrix is symmetrical and has trace elements equal to zero, it is best to compute only the upper or lower triangular portions. Next, one finds the smallest distance within the matrix and groups pairs of records represented by that distance (A,D). Using progressively larger distances, one adds records to existing groups or groups them with records which share the chosen distance (Table 1.2). The dendrogram for this nearest-neighbor cluster analysis is shown in Fig. 1.8. At this stage, one must subjectively choose a set of clusters that one believes to be useful. This is usually accomplished by drawing a horizontal line across the tree (Fig. 1.8). Although simple, this type of cluster analysis is not appropriate for all problems. For this reason, software packages support several algorithms. SAS (1985), for example, supports eleven. These include the k-nearest neighbor (KNN) algorithm and Ward's minimum variance method. The latter has connections with other statistical methods since it seeks to minimize within-cluster variance. On the basis of several reviews, Everitt (1993) recommends Ward's, group average and centroid methods as reasonably robust.

Table 1.1 Distance matrix for five records.

	A	B	C	D	E
A	0	3	5	1	2
B	3	0	9	6	1
C	5	9	0	8	3
D	1	6	8	0	2
E	2	1	3	2	0

Table 1.2 Nearest-neighbor groups at progressively larger distances. Par-entheses indicate hierarchical associations.

Distance	Clusters
1	(A,D),(B,E),C
2	((A,D),(B,E)),C
3	(((A,D),(B,E)),C)

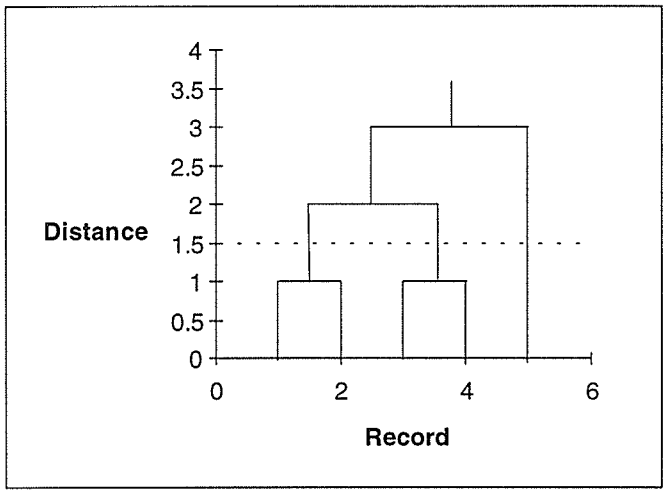


Fig. 1.8 Dendrogram for nearest neighbor hierarchical clustering.

Care should be taken in interpreting cluster diagrams for several reasons (Manley, 1994). First, these methods always find clusters, and will cluster patternless data. For this reason, it is wise to look for a consensus among several clustering algorithms applied to a given data set in order to gain confidence in the groupings. Statistical methods are also available to help assess the likelihood that a reasonable number or configuration of clusters has been chosen (SAS, 1985). Second, the choice of cutoff level on the dendrogram is subjective. Third, clusters that are not hyperspherical in multidimensional space may lead to poor representations. Fourth, the distance measure may be inappropriate for the data.

1.2.3 Artificial neural network (ANN) methods

Most discussions of ANNs begin by mentioning their biological analogs. Most multicellular organisms exhibit some form of communication between cells, be it chemical or electrical. Electrical communication is carried out in cells of an animal's nervous system. Invertebrates have relatively simple nervous systems, some like the nematode *Caenorhabditis elegans* (White et al., 1986) with less than 1000 nerve cells (neurons). This fact has allowed researchers to investigate functional details of actual

neural networks, and their components. Byrne (1987) for example reviews the physiological basis for associative learning in several invertebrates including the sea slug (*Aplysia* spp.), fruit fly (*Drosophila* spp.), and some vertebrates. This type of information, in turn, has inspired engineering research including the development of analog CMOS models for some neurons of *Aplysia* spp. (Card and Moore, 1990). Biology has also inspired the development of ANN theory.

In real and artificial neural networks, the basic element is the unit, neuron or node. Units receive inputs from several other units, perform some transformation on the input, and generate an output. Input connections are made across synapses. Individual inputs either stimulate or inhibit activity in the receiving unit to varying degrees. The amount and type of effect that an input has in an ANN is determined by a weight value. Units can be arranged in layers, with one layer providing some or all inputs to another layer. The manner in which units are interconnected to form a network is called the network architecture (Hertz et al., 1991).

One of the most basic unit models is often called the generalized McCulloch-Pitts neuron (Hertz et al., 1991):

$$y_i = f(\sum_k w_{ik} x_k + \mu_i). \qquad [40]$$

For one neuron, x_k are the k elements of an input vector, w_k the k weights and y the output. f(.) is called either the gain, activation, transfer or squashing function. μ is a constant input, called a bias or threshold. Depending on the architecture, different subscript formats may be used to generalize the equation for neurons arranged in layers, or for multiple input vectors. The squashing function has several common forms (Table 1.3), some of which are particularly useful since they are easy to differentiate.

Table 1.3 Common transfer functions for artificial neurons.

Name	Function
sign	$\mathrm{sgn}(x) = \begin{Bmatrix} 1 & if\ x \geq 0, \\ -1 & if\ x < 0. \end{Bmatrix}$
linear	$f(x) = x$
sigmoid	$f(x) = \frac{1}{1+\exp(-2\beta x)}$
hyperbolic tangent	$\tanh(\beta x) = \frac{e^{\beta x} - e^{-\beta x}}{e^{\beta x} + e^{-\beta x}}$

For simple problems, it is easy to pick weights that will generate the desired output values for a given input vector, especially if the transfer function is linear, or a step function.

One may also be able to find a unique mathematical solution for the weights and transfer function. In these cases, the problem reduces to finding the appropriate transformation matrix so that input vectors are mapped to the correct outputs. Less well defined problems require weights to be determined through an iterative technique. The process of adjusting weights is called learning, and there are many available learning algorithms. Learning is accomplished using a set of training data, while testing is usually accomplished with an independently chosen set of test data.

ANNs can be classified based on the learning algorithms that they use. The sections that follow describe some of these network types and the applications for which they are suitable.

1.2.3.1 Associative memories

Associative memories can reproduce an output pattern that resembles the input pattern. They can do so for a limited number of patterns that they have stored. For imperfect input patterns, they may be able to generalize by generating an output pattern that is at least a partial reconstruction of one stored pattern. If one was to store several bit-mapped images in an associative memory, and present noisy or incomplete versions as input, the memory should output the original image (Hertz et al., 1991). Unlike conventional content-addressable memories, these memories are associative on all bits of the stored word, rather than a restricted field of these bits. Such networks are useful for reproducing known patterns from noisy versions of them and also provide an instructive introduction to several neural network concepts.

An associative memory can be implemented by storing patterns in a look up table and outputting the pattern that most closely matches the input (Hertz et al., 1991). If the patterns are binary, one can use the Hamming distance to measure the dissimilarity. A neural network may also be constructed for the task by assembling a group of units:

$$y_i = \mathrm{sgn}(\sum_k w_{ik} x_k), \qquad [41]$$

and assuming that binary inputs and outputs are coded as ± 1, rather than zero and one. Although it seems reasonable to make output values a thresholded linear combination of the inputs, how does one determine what the weights are? One solution is to use some measure of the correlation among input patterns to set the weights; this is the 'generalized Hebb rule':

$$w_{ik} = \frac{1}{N} \sum_p x_i(patt_p) x_k(patt_p). \qquad [42]$$

The weight between the jth input and the ith unit is thus an average (or correlation) taken across all p stored patterns; N is the number of units. The reader may verify that this network works for a simple case where there are two units with input patterns of (1,-1)

and (-1,1). This model in which weights are updated individually is called the Hopfield model (Hertz et al., 1991).

Hertz et al. (1991) develop an in-depth theory of Hopfield networks using statistical mechanics. They employ Hopfield's concept of an energy landscape to visualize operation of a Hopfield associative memory. With the simple problem mentioned above, there are two possible outputs. These solutions can be visualized as two hollows (or minima) in a flat landscape. For a given input, the network ends up in one of these two minima. For more complicated networks and larger numbers of patterns, the landscape will contain many hollows of various depths. Some of these hollows will be false solutions or spurious network states. Development of this theory uncovered several facts which are pertinent to later discussions:

1. Errors occur because the act of storing many patterns not only generates desired memory states within the network, but may also generate spurious ones, due to interference between non-orthogonal patterns.

2. The probability of error increases with the number of patterns one attempts to store and decreases with the number of nodes.

3. The probability or error is higher for correlated or random inputs, than for totally independent inputs.

Hopfield further generalizated the model to include continuous valued units with sigmoid or hyperbolic tangent transfer functions. This brings us closer to the type of networks commonly used in pattern recognition tasks today.

1.2.3.2 Perceptrons

In this description of the perceptron information is drawn from Hertz et al. (1991), except where indicated. The simple perceptron is a single-layer feedforward neural network This means that a single layer of units responds to weighted inputs and produces a set of outputs. Weights are set by training the network with known input data for which the desired output is known. It is intended that the network is then able to produce the desired output for novel test data. We can specify units (without bias values) as:

$$y_i = f(\sum_k w_{ik} x_k), \qquad [43]$$

where the transfer function f() is usually non-linear. As in the Hopfield model, basic perceptrons use the sign function.

For notational simplicity, consider $\mathbf{o}_p$ to be the network output vector (predicted value) for a given input $\mathbf{x}_p$, and $\mathbf{y}_p$ the desired output (actual value) for the same input. As in statistics, it is desirable to minimize the differences between $\mathbf{y}_p$ and $\mathbf{o}_p$ (error or delta) by choosing appropriate weights. One chooses random weights initially and then adjusts them incrementally over time as input vectors are presented. If this is done in a way that

minimizes the error, we have an iterative algorithm which 'learns' the relationship between inputs and outputs by finding a set of appropriate weights:

$$w_{ik(new)} = w_{ik(old)} + \Delta w_{ik}. \qquad [44]$$

A simplified version of what is termed the perceptron learning rule, which is based on the Hebb rule (Hertz et al., 1991), is specified as follows:

$$\Delta w_{ik} = \eta(y_{ip} - o_{ip})x_{kp}, \qquad [45]$$

where the weight is between the kth input and the ith output, is adjusted after the pth training pattern has being presented. The size of the weight adjustment is limited by the learning rate parameter η, with $0<\eta<1$. Notice that as $\mathbf{y}_p$ and $\mathbf{o}_p$ converge, the weight changes approach zero.

Simple perceptrons can correctly discriminate between input patterns that are linearly separable (Minsky and Papert, 1988). This means that the input patterns are confined to regions that can be separated by a line in two dimensions, a plane in three dimensions or a hyperplane in higher dimensions. Unfortunately, their application is limited to simple problems.

An improvement on the perceptron can be made by using a transfer function that is differentiable. This addition permits other methods of weight adjustment to be applied. As was shown in the section on linear regression, the method of least mean squares (LMS) can be used to determine weights by defining a quadratic error function:

$$E[w] = \frac{1}{2}\sum_{i,p}(y_{ip} - o_{ip})^2 = \frac{1}{2}\sum_{i,p}(y_{ip} - \sum_{k} w_{ik}x_{kp})^2, \qquad [46]$$

and then finding the negative first derivative of E with respect to the weights; the constant of 1/2 is added for convenience. Adding the learning rate factor η permits one to generate the weight update rule called the delta rule, Adaline or Widrow-Hoff rule or the LMS rule:

$$\Delta w_{ik} = \eta\sum_{p}(y_{ip} - o_{ip})x_{kp} = \eta\delta_{ip}x_{kp}. \qquad [47]$$

This is the batch update form, in which the weights are calculated from all input patterns p in one step. It is possible to incrementally update the weights after each input is presented:

$$\Delta w_{ik} = \eta(y_{ip} - o_{ip})x_{kp} = \eta\delta_{ip}x_{kp}. \qquad [48]$$

This method is also called the stochastic update rule. Both methods, in practice, update the weights in a direction that shows the maximum downward slope within a localized area of the error landscape. It is worth noting that reliance of the rule on local gradients means that it can become trapped in a local minimum. This may not be the global solution that the user desires to reach. For this reason, the network is usually trained several times with different sets of random starting weights in order to reduce the probability of a sub-optimal solution being chosen (Haykin, 1994).

Multi-layer perceptrons have one or more 'hidden' layers between the input and the output layers of units. One of the most common learning algorithms is called back-propagation of errors (Hertz et al., 1991). In this document, multi-layer perceptrons which are trained with back-propagation are called MLP-BP networks. The following paragraphs describe the operation of a two-layer network (one hidden layer, one output layer).

For an input pattern p, the output of each hidden unit V_j is:

$$V_{jp} = g(\sum_k w_{jk} x_{kp}). \qquad [49]$$

Similarly the outputs o_i in the final layer with weights W_{ij} and with inputs from the hidden units have components:

$$o_{jp} = g(\sum_j W_{ij} V_{jp}) = g(\sum_j W_{ij}\, g(\sum_k w_{jk} x_{kp})). \qquad [50]$$

The normal quadratic error or cost function is given by:

$$E[w] = \frac{1}{2}\sum_{pi} (y_{ip} - o_{ip})^2 = \frac{1}{2}\sum_{pi} (y_{ip} - g(\sum_j W_{ij}\, g(\sum_k w_{jk} x_{kp})))^2. \qquad [51]$$

Applying the same techniques as one does for simple perceptrons, one can define the update rules for weights as:

$$\begin{aligned}
\delta_{jp} &= g'(\sum_k w_{jk} x_{kp}) \sum_i W_{ij} \delta_{ip}, \\
\delta_{ip} &= g'(\sum_j W_{ij} V_{jp})(y_{ip} - o_{ip}), \\
\Delta W_{ij} &= -\eta \frac{\partial E}{\partial W_{ij}} = \eta \sum_p \delta_{ip} V_{jp}, \\
\Delta w_{jk} &= -\eta \frac{\partial E}{\partial w_{jk}} = \eta \sum_p \delta_{jp} x_{kp}.
\end{aligned} \qquad [52]$$

It can be seen that the error y_{ip}-o_{ip} is propagated back first through the hidden-output weights W_{ij} by δ_{ip}, and then to w_{jk} by δ_{jp} ; this is where the name back-propagation of

errors originates. The unit transfer function is usually either a sigmoid or tanh function, both of which are easily differentiable.

To summarize a single training event, one begins with small random weights and presents an input to the network. The deltas (δ) can then be recursively calculated for all layers. Once the deltas are known, all weights are updated. This procedure is continued until all input records have been presented (usually in random order); presentation of all records to the network once is called an epoch. Many epochs are generally required to train the network, especially if the learning rate parameter η is small.

Although the basic MLP-BP suffers from problems like long training times and the presence of local minima, it can be successfully applied to a wider range of problems than the simple perceptron. A number of variations and improvements on the basic MLP-BP have been proposed since its inception. These have usually been aimed at addressing the weaknesses such as slow learning and spurious solutions that result from local minima in the cost (optimization) function. MLP-BP networks and their variations have been applied in many areas, including the recognition and generation of speech, analysis of protein structure, sonar target recognition, vehicle steering, image compression, games, character recognition and non-linear prediction and control.

1.2.3.3 Unsupervised Hebbian learning

Unsupervised Hebbian learning (Hertz et al. 1991) can be used to cluster observations in a data set, to extract patterns, or to reduce data dimensionality, many of the same tasks that multivariate statistical tools were designed to perform. One neural network algorithm in fact performs a PCA on input data. We discuss this first.

Consider a network of linear neurons with Hebbian learning:

$$\Delta w_i = \eta V x_i . \qquad [53]$$

With this simple model, weights grow without bound. In order to correct this deficiency, one can assume that the expected value of the weight change is zero.

$$E(\Delta w_i) = E(\eta V x_i) = E(\sum_j w_j x_j x_i) = E(\sum_j w_j C_{ij}) = \mathbf{C}\mathbf{w} = 0, \qquad [54]$$

where $\mathbf{C}$ resembles the correlation matrix of the input variables, but without real eigenvectors and with eigenvalues that are zero or positive. Unfortunately, this situation is unstable, because unless $\mathbf{w}$ is an eigenvector of $\mathbf{C}$ with eigenvalue zero, $\mathbf{w}$ grows without bound.

Oja (Hertz et al., 1991) solved this problem by using weight decay to contain the growth of **w**:

$$\Delta w_i = \eta V(x_i - V w_i). \qquad [55]$$

This ensured that **w** approached the eigenvector of **C** which had the maximum eigenvalue, and had unit length. Since **C** is not the covariance matrix, **w** does not originate at the centroid of the input space unless pre-processing is done to subtract the mean vector.

To find the first M principal components (using M output units) Oja's rule is modified (Sanger's rule):

$$\Delta w_{ij} = \eta V(x_i - \sum_{k=1}^{i} V_k w_{kj}). \qquad [56]$$

Such algorithms are of most use when one is attempting to generate principal components with real time data for which the underlying statistics are changing, or when one is dealing with very large data sets. In these cases, the statistical approach described earlier becomes prohibitive.

There are other unsupervised feature extracting networks based on Hebbian learning. Linsker's Hebbian network has several layers in which units are connected to a limited number of units of the preceding layer (Hertz et al., 1991). These limited 'receptive fields', combined with Linsker's learning rule, permitted the network to learn feature detecting units similar to cells found in the mammalian eye. Other unsupervised techniques are outlined in the following section.

1.2.3.4 Unsupervised competitive learning

For completeness, this section was included to provide a brief outline of some other unsupervised Hebbian networks (with weight decay) that are available; the information is based on Hertz et al. (1991). Such networks find applications in clustering, vector quantization (coding of input records into classes), optimization problems, function approximation and signal analysis. The basic concept is that only the output with the largest net input (the winner) is active in what is called hard competitive learning. Output units are connected to input units via the usual weights, but outputs may also connected to each other by inhibitory weights (hard competitive learning) or not (soft competitive learning).

Simple competitive learning operates by updating the weight vector of the winning unit for a given input. Weights for other units are not modified. The process is such that an input causes the weight vector for the winning unit to migrate toward it. The standard competitive learning rule accomplishes this:

$$\Delta w_{i^*j} = \eta(x_{jp} - w_{i^*j}). \qquad [57]$$

This means that the weights w_{i^*} for the winning unit i* are updated given input pattern p. Research in the area of competitive learning has generated many new techniques that are based on this basic concept. Some of these are Grossberg's adaptive resonance theory (ART) and Kohonen's self organized feature mapping (SOFM) algorithms. Variations include learning vector quantization (LVQ), counterpropagation and radial basis function (RBF) networks which combine both unsupervised and supervised operations. As with cluster analysis, which has a number of available algorithms, each type of unsupervised neural network has its own advantages, disadvantages and appropriate applications.

1.2.4 Comparison of methods

A paper by Sarle (1994) sparked my interest in the relative merits of artificial neural networks (ANNs) and traditional statistical techniques when applied to pattern recognition. It became apparent during this study that the techniques complemented one another. A subsequent literature review uncovered several papers that compared and contrasted artificial neural networks (ANNs) with more traditional statistical analyses. Reviews on the topic have been written by statisticians such as Ripley(1994) and Cheng and Titterington (1994). These two reviews were presented as seminars and included written comments from a number of other researchers. This literature provides valuable insight into the existing state of the two research areas, and the utility of the techniques.

Amari (1994) recognized three classes of ANN research:

1. Biologically-related research that involves physiology and molecular biology.

2. Theoretical research into mathematical properties of ANNs.

3. Engineering applications such as pattern recognition and parallel computation.

Statisticians have interest in the last two areas. I focus on the last area since my interest is in pattern recognition. It turns out that ANNs useful for pattern recognition are also those that have attracted the attention of statisticians. It is interesting that Amari chose the term 'engineering applications'. This indicates that ANNs have mainly been used within the engineering community.

Breiman's (1994) perspective is that ANNs initially attracted unnecessary hype. The connotations of words like learning, brain and artificial intelligence stimulated excitement about this emerging field, but may also have been misleading. Breiman now sees ANNs as playing a useful scientific role, and draws attention to those researchers currently using them for pattern recognition. Researchers in this community use ANNs, hidden Markov models, and a variety of multivariate statistical methods to achieve their goals. He characterizes them as pragmatists who use whatever technique yields the best performance for the application of interest. There are also statistical and neural network communities. Statisticians focus on underlying models and unified methods while ANN

researchers focus on accuracy and tend to customize their ANNs to their application. Those who use both techniques have discovered that appropriately chosen ANNs and statistical methods may perform equally well on a given problem or they may differ (Ripley, 1994; Cheng and Titterington, 1994; Breiman,1994).

Ripley (1994) compares the performance of multilayer perceptrons (several architectures) trained with back-propagation of errors (MLP-BP) and some statistical methods for discrimination. In one case, neural networks performed as well as most of the statistical techniques. In another, statistical techniques worked better. It is particularly interesting to note that this latter case was a reanalysis of the classic sonar data of Gorman and Sejnowski (1988). Papers are now being published in which researchers include both ANNs and statistics in their analysis toolbox. Anderer et al. (1994) has applied MLP-BP, *z*-statistics and linear discriminant analysis to the task of discriminating between normal and demented patients based on EEG patterns. Although they concluded that the ANN performed better than the two statistical methods, they did not exhaust the range of applicable ANNs or statistical methods. Similarly Benediktsson et al. (1990) compare MLP-BP networks to a Bayesian classifier for vegetation classification from remote sensing data. Their conclusion is that the Bayesian classifier outperforms the ANN on novel data because it contains an underlying model. It is unclear, however if this was a fair comparison because learning in the ANN was stopped based on the value of the error gradient with training data. With either technique, the key to success has been appropriate preprocessing.

Table 1.4 A comparison of multivariate statistics and artificial neural networks for pattern recognition.

Characteristic	Multivariate Statistics	Artificial Neural Networks
Computation required	Low (especially if not iterative)	Can be very high (iterative)
Ability to evaluate an underlying or a priori model	High (many methods require the specification of a model)	Low (an underlying model is difficult to incorporate)
Practicality with very large numbers of records	Lower (matrices of large dimension are difficult to process)	Higher (given enough time, almost any data set can be processed)
Ease of application	Simple to complex	Generally simple
Ability to determine importance of input variable	High (can easily assess the relative contributions of variables to the predictive	Low-high (must examine relative magnitude weights to assess the relative contributions

	power of a model)	of variables)
Ability to quantify performance	Generally high when model based approaches allow probabilities to be assigned	More difficult
Ability to assess appropriateness of method	High (assumptions are stated)	Low
Possibility of a unique (global) solution	Guaranteed with some methods (iterative methods may yield sub optimal solutions)	Not guaranteed (local minima in the optimization function can occur)
Sensitivity to choice of parameters	Predictable	Relatively predictable
Requirement for validation	Generally low	High
Ease of choosing most appropriate method	High (there are prescribed guidelines for achieving a given goal)	Low (it is usually necessary to tune the network architecture and training methods)
Performance with non-linearity	Poor to excellent (depends on the method)	Generally good

Table 1.4 was constructed to provide a qualitative comparison of multivariate statistics and neural networks based on both literature and personal experience; Fig. 1.7 (at the beginning of the statistical methods section) is relevant since it suggests that ANNs can be used for specific research goals.

One can also derive some guidelines for the application of neural networks and statistics to pattern recognition problems from information presented in this section. These are presented in the following list of 'traps' to be avoided when applying these techniques.

Traps for users of statistics:

1. Assuming that neural networks are not appropriate.

2. Abandoning a data set when an appropriate underlying model cannot be found.

3. Failure to carefully evaluate the validity of assumptions based on the data set.

4. Spending too much time modeling data (Breiman, 1994), when what is desired is something that works well for the intended purpose.

Traps for users of artificial neural networks:

1. Assuming that statistical methods are not appropriate.

2. When there is an underlying model, ignoring it in the hope that the network will discover it.

3. Failure to assess the optimality in terms of training time and network architecture.

4. Failure to perform validation with a test set.

5. Termination of training on the basis of training data performance only.

Traps for users of both:

1. Assuming that an ANN or a statistical method is a 'black-box' problem solver. No one technique is appropriate for every problem.

2. Using data that is not representative of the population; this includes the problem of having a sample of insufficient size.

3. Not spending the time to find the most appropriate way to process data (data processing and transformation, preprocessing) so that the method performs well.

4. Choosing an inappropriate statistical method or ANN type for the problem.

5. A belief that the model you have proposed is somehow 'the true model'.

6. Belief that extrapolation is made without risk.

1.3 Analysis of bird sound

It is instructive to examine the traditional means used by biologists to analyze and describe bird sound. The earliest work consisted of qualitative and often subjective descriptions of the sounds birds made. Judgment was used to determine whether or not songs (or their components) were similar. Recent work may involve some subjective criteria for distinguishing sounds, but generally quantitative and objective techniques are preferred.

This section covers several topics. First, procedures for putting a sound in a form from which measurements can be taken are described. Next, the types of features measured and the nature of variation in bird sounds is discussed. Lastly, a brief summary is provided of statistical analysis techniques that have been applied to bird sounds.

1.3.1 Preparation for analysis

As with speech, bird sounds can be examined on a frequency-time plane. Analysis usually begins by obtaining a frequency-time plot of a recorded bird sound. The sonagraph was the instrument traditionally used for this purpose (Aubin, 1992). It consisted of a bank of bandpass filters whose thesholded outputs controlled the application of ink to a strip chart in real time. Narrow and wide band options were usually available. Blackened areas on the chart indicated that certain energy was present at a given frequency-time coordinate. Measurements of selected temporal and spectral parameters can then be taken manually from sonagraph plots.

Presently, software and electronic hardware permit the same measurements to be made using a graphical user interface (McGregor, 1991). Tools designed for speech analysis are commonly used for this purpose. Analysis methods include FFT, PSD and LPC techniques described earlier. In addition, biologists have considered zero-crossing analysis, autocorrelation, cepstral analysis, Wigner-Ville transforms and wavelet transforms (Aubin, 1992).

The degree of objectivity and depth of quantitative analysis used by biologists in bird sound research varies in accordance with the variety of their study goals. In this research, I was interested in two facets, which can be stated in the form of questions. First, does the literature indicate features that are useful in discriminating between species? Second, how variable are the measured features? Lastly, what are there analytical tools used by biologists for bird sound analysis, and are they of use in automated recognition?

1.3.2 Features for discrimination

Becker (1982), in his review, provided a good description of the sound characteristics birds actually use for recognizing members of their own species. Table 1.5 indicates that both temporal and spectral characteristics are important in bird communication.

Table 1.5 Importance of certain sound characteristics for species recognition in birds .

Characteristic	Important?
Sound length	Not by itself (but see syntax)
Amplitude (emphasis)	No; too easily affected by the environment.
Interelement interval	Yes; responses are affected by rhythm.
Frequency range, tonality	Yes; responses may be affected by filtering.
Syntax	Yes; repetition, and ordering of elements may affect response.
Element structure:	
Frequency modulation	Yes; many species are sensitive to this.
Amplitude modulation	No; fine variations are lost in transmission.
Transients	Generally not, although some species require elements to begin after at least a short silent period.
Tonality	Sometimes; some species are affected by addition or removal of harmonics.

Summarized from Becker (1982).

Table 1.5 suggests that tonality and frequency changes are important spectral characteristics, while interelement intervals and syntax are important temporal cues.

Becker (1982) reviewed evidence that different bird species respond to different song features. The Goldcrest, for example, is sensitive to the change in frequency between consecutive elements of a song; reducing the frequency difference reduces the response of males to artificially modified test songs. In contrast, the related Firecrest (which is found in the same habitat) responds most strongly to the test songs that contain the least frequency change. Using similar experiments, it has been shown that constant frequency portions of elements (Willow Tit), frequency sweeps (Marsh Tit), element syntax (Winter Wren), repetition of sounds (Brown Thrasher) and temporal pattern (Song Sparrow) are used by birds to discriminate between songs of their own species and those of other similar sounding species. Such research also indicates that birds are capable of discriminating between the songs of two species, even when humans (with technical aids) are unable to do so.

Several researchers discovered that invariant or stereotyped features (those which vary little among populations or species) exist within songs of a given species (e.g. Smith *et al.*, 1980). This has led to the suggestion that such features should be used by birds in recognition of their own species (Becker, 1982). Nelson (1989), however demonstrated that this is not always the case. Using 14 measured features from Field and Chipping Sparrow song, and canonical discriminant analysis, Nelson showed that variability of feature variables within species did not predict the variable which best discriminated between species.

Although it seems contradictory, the information presented so far in this section provides guidance toward development of an automatic recognition system. First, different bird species use different cues to recognize members of their own species. Second, it is clear that features useful for discriminating between members of one group of species may not be useful in discriminating between those of a different group. Third, features that are invariant within a given species' song may not be the best ones to use for inter-species discrimination. The researcher is thus left with a choice between two approaches for automatic recognition:

a) extract features that are known to be of general importance and use a single classifier, or

b) extract features that are known to be important for discrimination within limited sub-groups of species and use different classifiers (specialists) for each sub-group; final identification would require the combining of results from specialist classifiers.

For preliminary investigations, such as this one, the first approach is simpler and less time consuming to implement.

Having a rough idea about the features birds extract from their own sounds, it is now instructive to examine methods used by biologists in the study of bird sound. Researchers do not use a standardized procedure for describing bird song temporal features of bird song (Thompson et al., 1994). When one examines the literature, it is also apparent that the choice of spectral measurements varies. Generally, manual measurements are made from frequency-time plots, with features being chosen somewhat subjectively. A partial list of measurements that have been used includes: highest energy frequency within an element (Weary et al., 1991); elements per song, element duration, interelement interval, song duration, maximum and minimum frequency, frequency range (Rich, 1981); duration and slope of FM sweep (Date et al., 1991); time and frequency coordinates of frequency inflections (Payne and Budde, 1979); an 'element' is a burst of sound that is separated from other such bursts by a perceptible pause (inter-element interval). Figs. 1.9 and 1.10 illustrate some of these features using a Fox Sparrow song.

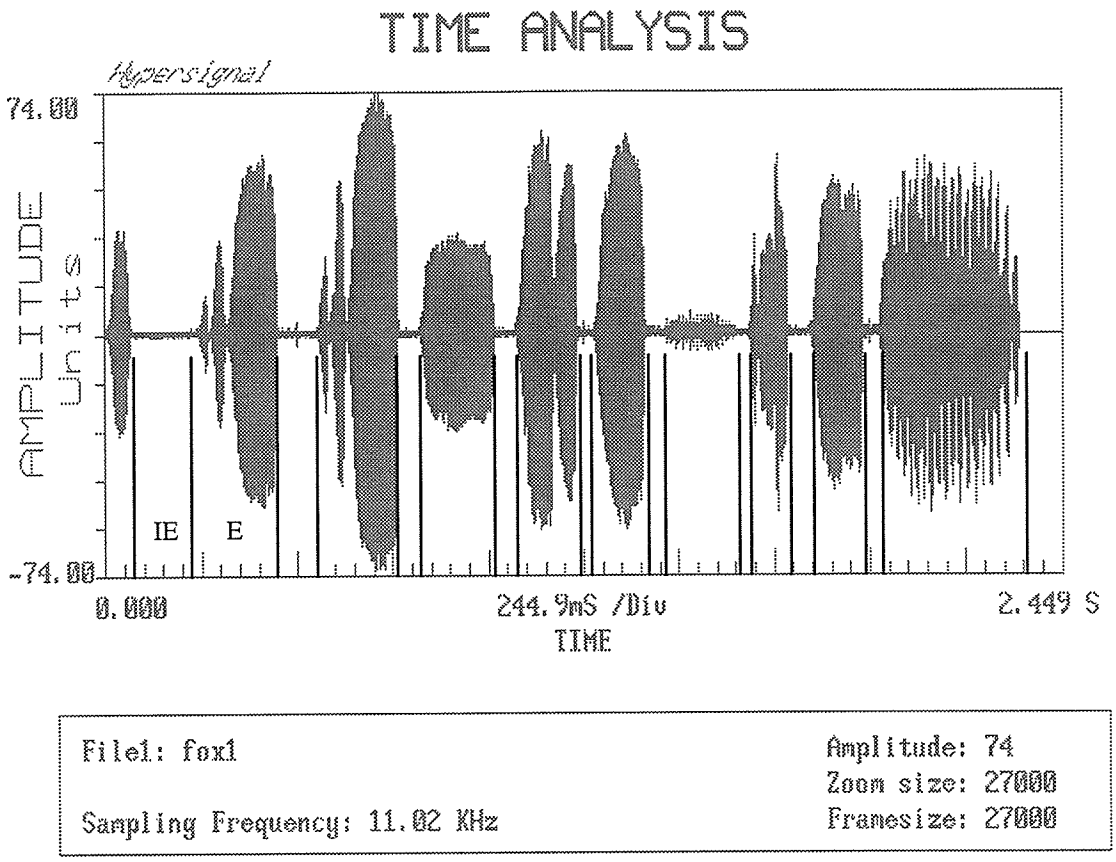


Fig. 1.9 Time domain representation of a Fox Sparrow song. Vertical lines indicate boundaries between elements (E) and inter-element intervals (IE). For clarity, only one example of each feature is labeled.

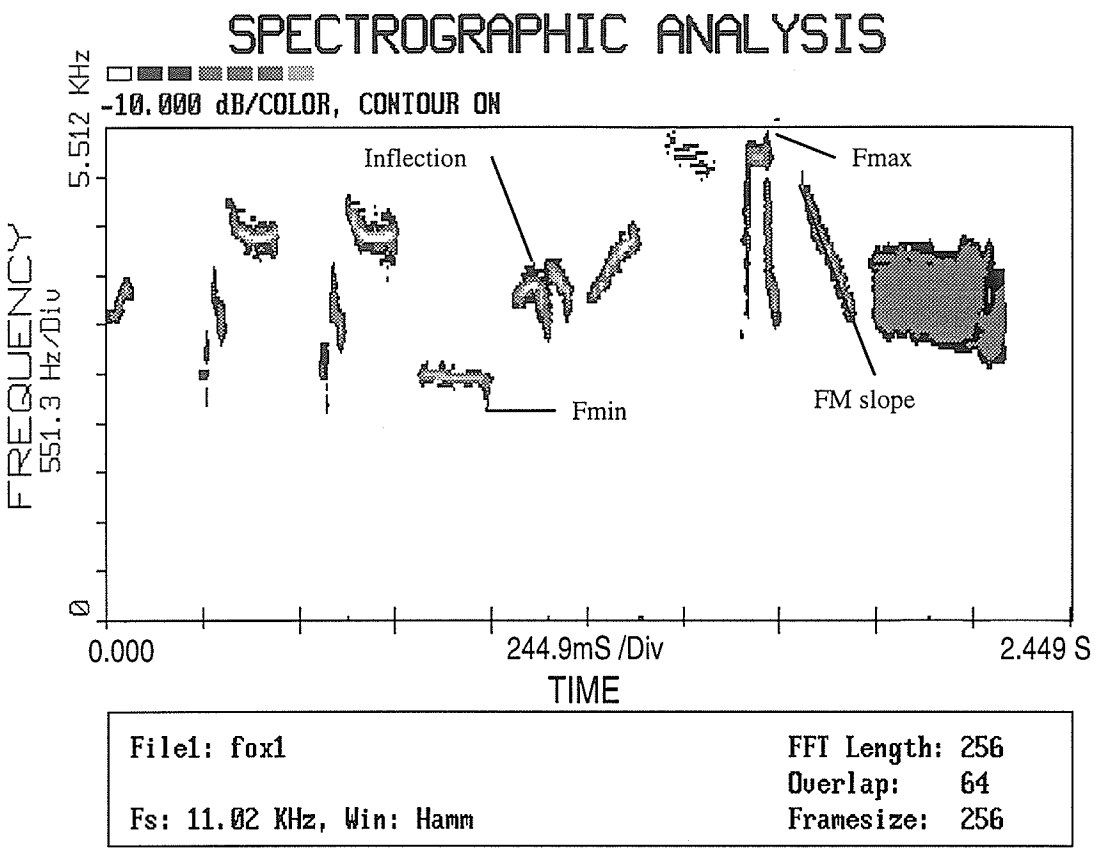


Fig. 1.10 Frequency domain representation of a Fox Sparrow song. Maximum frequency (Fmax), minimum frequency (Fmin), a frequency inflection and the slope of an FM sweep are indicated. The FFT was performed with a Hamming window.

Although most of the parameters mentioned in this section have the potential to be measured automatically, there are few examples in the literature. Okanoya and Kimura (1993) designed a simple automatic call detector that uses zero-crossing rates and song length to discriminate between two species of finch. Evans and Evans (1994) developed an automated pattern matching technique. They parse a digitized song into elements (sounds separated by at least 200msec of silence). Each element is first compared to all elements in a library on the basis of length. If any library elements match (length differs by <10%) a form of cross correlation is performed. These comparisons did not consider spectral characteristics.

1.3.3 Variability in song

In order to appreciate the difficulty in devising a technique for automatic identification of birds by sound, one must examine the variability in bird sounds. The songs of six bird species were used in this study: Song Sparrow (*Melospiza melodia*), Fox Sparrow (*Passerella iliaca*), Marsh Wren (*Cistothorus palustris*), Sedge Wren (*Cistothorus platensis*), Yellow Warbler (*Dendroica petechia*) and Red-winged Blackbird (*Agelaius phoeniceus*); nomenclature follows Robbins et al. (1983). Fortunately, songs of each have

been studied at one time or another. In the paragraphs that follow, the nature of these songs and their variability are briefly described; the literature cited is only a sample of that which is available.

Bankwitz and Thompson (1979) examined 745 Yellow Warbler songs recorded in Michigan and concluded that there were four accented ending (AE) song types and 60 unaccented ending (UE) song types. Each song consisted of several elements of which there were 40 unique types. Even with this variation, different populations (groups of individuals) exhibited much overlap in their use of elements. The frequency of AE and UE songs did vary with season. In a more detailed study, Weary et al. (1994) found that individual males may either repeat one song, or run through a repertoire of several songs, depending on the behavioral context in which they were singing.

Smith et al.(1980) analyzed over 500 Red-winged Blackbird songs from six males and concluded that some parameters of the song were strongly stereotyped, while others were not. The song sounds like "conc-a-ree", with the "ree" portion being a modulated trill (rapid series of sound bursts). The rate at which the trill was modulated varied little. This variation however was comparatively large when compared to stereotyped parameters of songs like that of the Chipping Sparrow.

Fox Sparrow songs were extensively studied by Martin (1977, 1979, 1990) and were categorized into five general types. Individuals sang a repertoire of songs in a particular order repeatedly, and even maintained this order over several years. This pattern was consistent enough to be accounted for by a first-order Markov model. Individuals used similar songs, but sang them in different orders. The 49 types of elements used to make up songs tended be used consistently over several populations.

In the Song Sparrow, Baker (1983) reported that songs were composed of 72 element types, and that individuals used from 8-18 different song types in their repertoires. Martin (1990) also cites work of Mulligan which indicates that Song Sparrow song is very variable, even within one individual.

Kroodsma and Verner (1978, 1987) have examined Marsh Wren songs recorded in a number of geographic areas. From recordings of 18 individuals, males had from 36-195 song types. However, the duration of two parts of the song, termed the trill and the song, varied less within populations than between populations.

Although Sedge Wrens and Marsh Wrens share similar habitats, Kroodsma and Verner (1978) found that Sedge Wrens (in contrast with Marsh Wrens) in the same population had more variable song types and larger repertoire sizes. Sedge Wren songs were found to have a consistent introduction (4 types), followed by a variable trill.

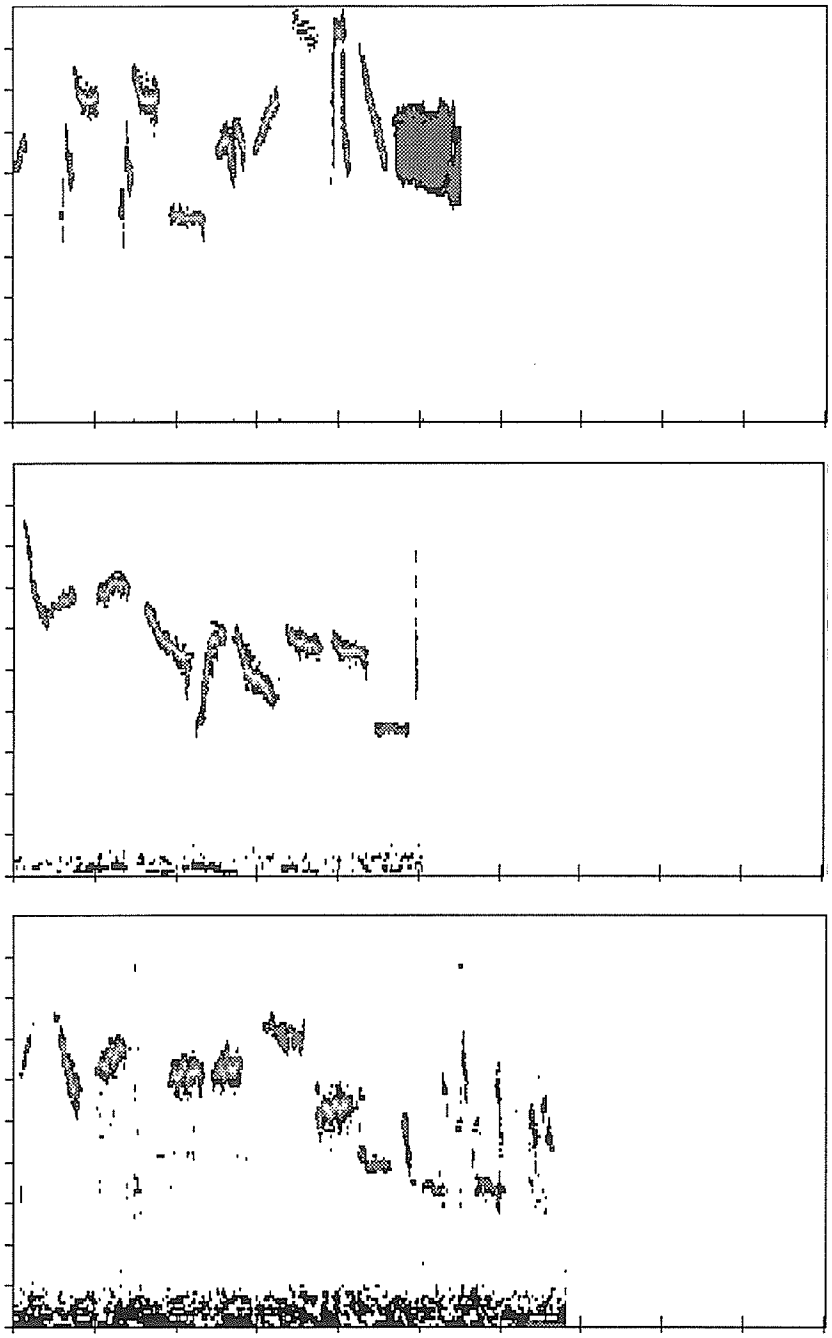


Fig. 1.11 Spectrograms of three Fox Sparrow songs. The time range is 0 - 4 seconds, the frequency range 0 - 5.5125kHz.

The information presented for these six species suggests that bird songs can vary a great deal within a species. A small example of this is illustrated in Fig. 1.11 for Fox Sparrows; here, the length, number of elements and frequency characteristics all vary. To draw a link to research on human speech, the degree of variation observed within a bird species probably exceeds that observed in speaker-independent word recognition.

Song variation may be detectable between geographic regions, populations or even individuals. Furthermore, the behavioral context or time of year in which the bird sings may also affect the song sung. Again, this has parallels with speech recognition, where the one encounters differences due to accent, dialect, and an individual's sex, age or state of health. In speech recognition, similar sounding words are confused; some bird species have similar sounding songs (eg. Rose-breasted Grosbeak, American Robin, Scarlet Tanager).

Any method of automatic identification must be more sensitive to interspecific variation than to intraspecific variation. Since it is reasonable to assume that some features of even the most variable songs are stereotyped, the task of automatic identification only becomes possible if the system is capable of extracting such stereotyped features.

1.3.4 Statistical analysis

As in the previous section, a review of the biological literature on bird sound analysis reveals the use of several statistical analysis techniques. Some work is summarized here in a brief overview of methods and their applications.

PCA has been used to study song variation in Solitary Vireos (Martindale, 1980), Rufous-collared Sparrows (Handford and Lougheed, 1991) and Acadian Flycatchers (Payne and Budde, 1979). Clustering has provided information of song variation in Fox Sparrows (Martin, 1979), Solitary Vireos (Martindale, 1980) and Prothontary Warblers (Bryan et al., 1987). Discriminant analysis was used by (Rich, 1981) to verify the existence of dialects among Sage Sparrow populations and by Lemon et al. (1983) to show that song features could be used effectively to discriminate between warblers. Canonical discriminant analysis was applied in a study of song variation among two species of chickadees (Robbins et al. 1986). A Markov chain model effectively modeled song sequences in Fox Sparrows (Martin, 1990). Sparling (1978) applies a wide variety of methods in a tutorial illustration of the applicability of multivariate statistics to the analysis of bird sounds. ANOVA is often used where simple treatment comparisons are required.

This literature indicates that the statistical analysis techniques described earlier have been widely used in bird sound research. In at least one instance, statistics was applied to the task of species recognition. Integrating information from biology, computer engineering, and statistics provides a firm basis for development of an automatic species identification system.

1.4 Justification and goals of research

This research originally began with the idea of applying digital signal processing and ANNs to the problem of identifying birds to species based on the sounds they make. It soon became apparent that there was a practical interest in an automatic recognition system. Given that vocalization and audition processes used for communication in birds are similar to those used for speech, it seemed reasonable to apply traditional speech analysis and recognition techniques to the problem.

As with speech data, the volume of data generated by sampling the time domain waveform of a bird sound is large. This made it necessary to use pre-processing to extract relevant features prior to the pattern recognition step. Biologically inspired methods for pre-processing of speech such as LPC analysis were thought to be a good starting point for dimensionality reduction of data.

Biology also suggests that complex non-linear transformations are used in the brain to identify sounds. Lippmann's (1989) review of neural network architectures for speech recognition suggested several techniques for classifying / recognizing static speech windows; data were usually vectors of spectral parameters extracted from the samples in the window. This was analogous to the type of processing that occurred in the cochlea. In addition, information indicating vertebrates had neurons that were delay sensitive suggested that temporal parameters were important in animal hearing. Thus, the research was directed toward using DSP techniques to extract temporal and spectral parameters from sounds. The ultimate goal was to fully automate parameter extraction to avoid the subjectivity inherent in manual determinations. Next, pattern recognition was required.

ANNs initially appeared to be the most applicable technique for identifying the bird to species based on its sound because they had the ability to learn complex non-linear transformations. Further literature research led to the thought that multivariate statistical techniques could be used for recognition. Since statistical methods also provide efficient means for selecting subsets of variables for analysis, it would be possible to select a subset of the extracted parameters for analysis. Such a reduction in data dimensionality would be of benefit for both statistical and ANN recognition techniques.

This line of thought formed the basis of the literature review and is reflected in the two research chapters. Chapter 2 presents the initial implementation of an identification system, which utilized LPC analysis for spectral estimation, and simple temporal parameters. A supervised ANN was used for pattern recognition. In Chapter 3, another method of parameter extraction is developed which was inspired by information on key features used by birds in species identification. ANNs and multivariate statistics are applied in pattern recognition.

2. SPECIES IDENTIFICATION - METHOD ONE

2.1 Introduction

Since birds generate sound in a similar manner to the way in which humans generate speech, the source-filter model of speech production may be applied to the analysis of bird song. If one assumes that vocalizations of both birds and humans can be modeled by source-filter interactions, one may then characterize these vocalizations by extracting filter coefficients from the time domain waveforms. Linear predictive coding (LPC) fits the impulse response of an all-pole filter to such a waveform. Performing an FFT on time domain coefficients of the whitening filter $P(z)$ then provides a smoothed representation of the spectral envelope. Such an analysis provides a good starting point for the analysis of bird songs.

As with isolated word recognition, there are several difficulties to overcome before automated birdsong recognition can be attempted. Variability of vocalizations due to dialect and individual differences obscures word and song identity. The temporal nature of sound also creates both dimensionality and context problems.

To deal with variability, and the lack of theoretical information on the structure of bird songs, an artificial neural network was chosen for the recognition task. The ability of an ANN to perform non-linear mapping without requiring parametric assumptions has been successfully employed for pattern classification and speech recognition (Lippmann, 1989b),(Lippman, 1989a).

Using LPC coefficients automatically reduces the dimensionality of input vectors. Taking the FFT of the LPC coefficients also provides a means of data reduction; phase information can be discarded since human speech processing appears to be relatively unaffected by phase information (Denes and Elliot, 1993). Thus use of LPC analysis and removal of phase information results in useful compression of sampled data.

Temporal context can be incorporated into a word recognition system that utilizes neural networks in several ways (Lippmann, 1989b). The most obvious method is to present a network with the entire spectrogram as an input vector (using FFT coefficients for all frames in a vocalization). Such a system has been used effectively to model the target recognition that bats appear to perform (Dror et al., 1995). This technique, however, has a high computational cost since large numbers of weights must be updated in training. Others have used time-delay neural networks, in which both present and a few past frames provide input parameters to the network. Waibel used this technique together with the concept of large modular networks to recognize consonants (Waibel, 1989). Recurrent networks that have context feedback units have also found applications (Hanes et al., 1994).

One problem with word recognition is that it is computationally intensive. With neural networks, a way to reduce computation is to limit the number of nodes in the network. In addition, proper coding of input data is critical (Hertz et al., 1991). For this reason, an

attempt was made to implicitly encode temporal context through pre-processing, rather than expecting the network to discover temporal context from a high-dimensional input. The goal of this section is to demonstrate that a simple neural network is able to identify a bird species based upon its song, a task which is referred to as species recognition.

2.2 Methods

2.2.1 Sampling

Data for the study were extracted from audio tapes and compact disks that were intended for use by people wishing to identify species by song (Brigham, no date; Borror, 1970; Borror, 1967; Elliot and Mack, 1990; Walton and Lawson, 1989; Kellogg et al., 1975). 133 songs were chosen so that each originated from a different individual. Recordings varied in quality, with some having discernible tape hiss or in some cases the presence of other songs in the background. There was little detectable reverberation. Six species of bird native to Manitoba were chosen: Song Sparrow (SON), Fox Sparrow (FOX), Marsh Wren (MWR), Sedge Wren (SWR), Yellow Warbler (YLW) and Red-winged Blackbird (RWB) (Robbins et al., 1983). They were chosen to provide a representation of both long and short songs. Songs from these species were digitized using a PC sound card (11.025kHz sampling rate, 8 bit resolution). Digitization was performed with AGC, and levels were adjusted to give maximum amplitude without clipping. This was done to remove variation in the FFT magnitudes due to amplitude variation.

2.2.2 Preprocessing

Data was left-justified by hand and manipulated by the software package Hypersignal Plus, which preprocessed the data in several steps:

1. Framing -- a non-overlapping Hamming window, 512 samples (approx. 46 msec) wide was applied. This is similar to frame widths used for speech recognition, which are normally in the order of 10-30 msec (Lippmann, 1989b). Note that longer songs produced more frames.

2. LPC -- 16 time domain coefficients for a 15th order LPC filter were generated for each frame.

3. FFT -- a fast Fourier transform of the 16 whitening filter coefficients was used to produce 9 unique spectral magnitudes.

This procedure was repeated with a 2048 sample window for all songs. The resulting data were exported from Hypersignal, with each song being represented by a number of records, each containing nine spectral coefficients. Further processing of the data was required before it could be used for training the neural network. Initial investigation revealed to us that the overall length of a bird's song was an important cue in species identification. Adding such a variable often helps a network learn from hints (Haykin, 1994). For this reason, a variable was included to represent the length of a song. Its value was the same for each record within a given song. Spectral and time variables were

normalized to a mean of zero and a standard deviation of one. Variables were squashed using a logistic function with a gain of unity.

Three types of data set were created. Two had 10 variable records, derived from either the 512 sample window (9) or from the 2048 sample window (9) and the song length (1); the song length variable was merged with variables from each frame of a given song; merging involved repeating data for length and 2048 sample variables to create the 19 variable records. The third data set included all 19 variables; records from the 2048 sample window were merged in a manner similar to that used for song length. This process is summarized in Fig. 2.1.

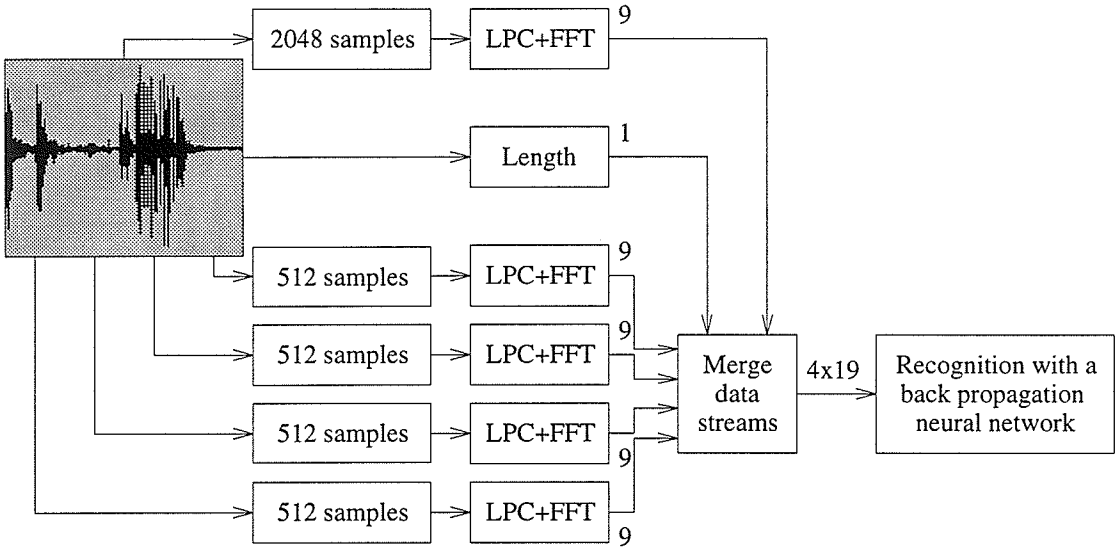


Fig. 2.1 Schematic example of pre-processing using combined 2048 and 512 sample windows for a 2048 sample sound. Numbers next to lines indicate the number of variables at each step; four 19 variable records are used as input for the recognition step.

2.2.3 Identification

In this study 'vanilla' back-propagation, without momentum or higher order derivative information was chosen (Rumelhart, 1986) as the learning model. Training was accomplished using the PDP (McLelland and Rumelhart, 1989) back propagation algorithm. Some experimentation suggested that a network of 10 or 19 inputs, 12 hidden nodes and 6 outputs would work well. The learning rate was set to 0.2. Target values of 0, 1 were changed to 0.2 and 0.8 to accelerate learning.

Songs were divided between test and training data sets in random order (records within songs were not randomized at this stage). The proportion of data used for testing was 25% in all runs. To evaluate performance, 10 training and 10 test sets were generated for each of the three data set types, and the network was trained with new initial weights each time. The 10 test and 10 training sets were generated in the same order for all three data sets in order to facilitate comparisons. Preliminary training runs of 10000 epochs (passes

through the whole training set) suggested that a run length of 1500 epochs was sufficient for the mean (over n = 10 trials) sum-of-squares error (MSE) to stabilize.

Final weights were read by a C program that computed the six output values for each of the test set records and an error sum-of-squares (for actual vs. expected outputs). For each song, any activation > 0.6 was counted for the 6 species outputs. As illustrated in Table 2.1, this technique was designed to yield consistent and somewhat conservative results; it was relatively insensitive to cases where several outputs were activated for a given frame. Frame seven illustrates the rear event where two outputs were activated at a level greater than 0.6. To establish the identity of the bird, the species class with the largest count was considered to be the winning class. If two classes were tied for the maximum count, classification was considered to be incorrect.

Table 2.1 Example of neural network output for some 512 sample frames of Yellow Warbler song number 17. Values are node activations for a given frame, with values over 0.6 being in bold-face type; such activations are counted in the last row.

Frame	RWB	YLW	MWR	SWR	FOX	SON
1	0.24	0.11	0.29	0.20	0.18	**0.80**
2	0.20	0.17	0.39	0.24	0.24	0.47
3	0.39	0.03	**0.73**	0.18	0.29	0.22
4	0.25	0.10	0.29	0.20	0.18	**0.81**
5	0.26	0.14	0.37	0.26	0.26	0.41
6	0.25	0.11	0.29	0.20	0.18	**0.78**
7	0.19	0.02	**0.61**	0.25	0.18	**0.77**
Count	0	0	2	0	0	4

2.3 Results

The mean MSE during training is largest for the 512 sample window data, smaller for the 2048 window and smallest for the combined data set (Fig. 2.2). Even though the differences in MSE appear to be small, they are significant over the duration of the training cycle. Two-tailed t-tests (Bhattacharyya and Johnson, 1977) comparing the mean MSE (between 512 and 2048, 2048 and 'BOTH', n = 10) at epoch 1500 indicate p-values << 0.0005. This means that there was only a small chance that the mean final MSE values were different due to chance.

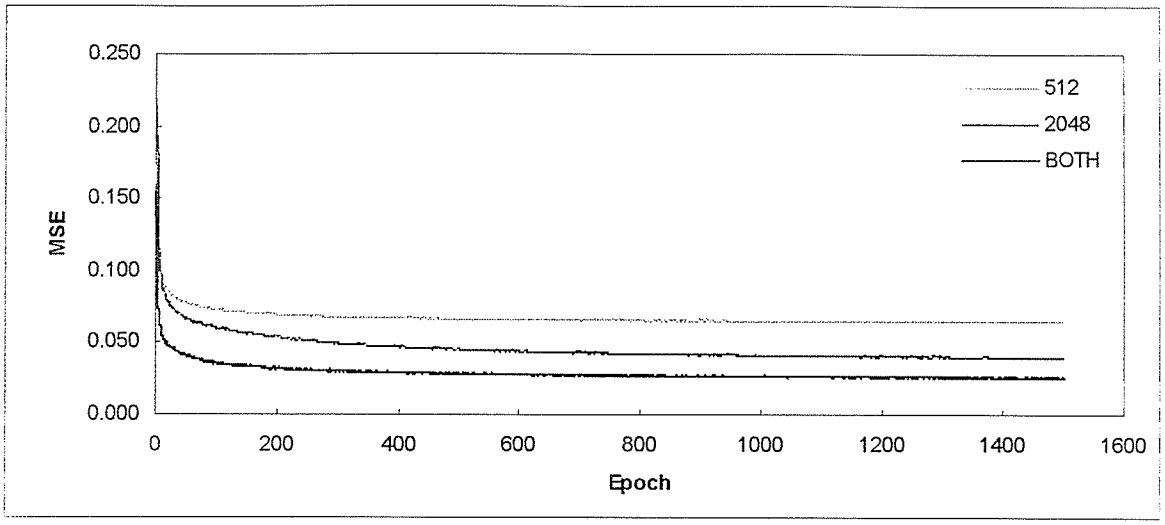


Fig. 2.2 Learning curves for a back-propagation network with 512 sample, 2048 sample and combined window data sets. MSE is the mean sum-of-squares error over all training records, and averaged over 10 trials.

Fig. 2.3 shows that the network was able to recognize the six species by their song. Overall performance ranged from 91% to 93% correct identification. Except for the FOX and RWB, mean performance increased as the input was somewhat higher for 2048 and combined data sets than for the 512 data sets.

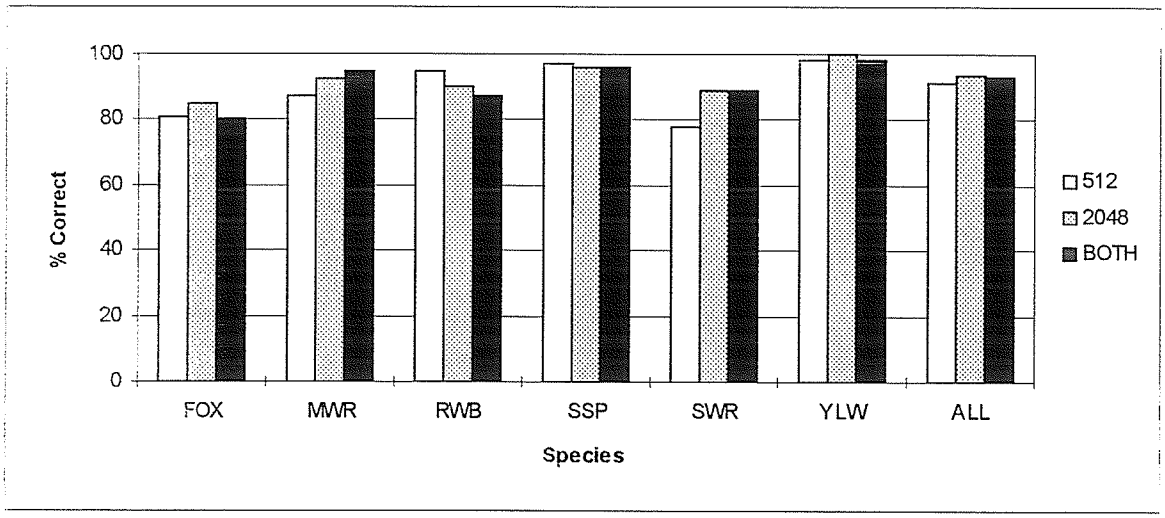


Fig. 2.3 Species recognition performance for three neural networks. The category 'ALL' represents performance over all species.

Table 2.2 shows that there was considerable overlap in song length. Song lengths of YLW, MWR and SON had bimodal frequency distributions. When raw results from the runs were examined, it was noticed that some songs were consistently misidentified. In two cases, a song was either atypically long or short, and was misidentified regardless of the data set used. Other songs were misidentified with 512 and 2048 window data sets, but were correctly identified with the combined data set.

Table 2.2 Length of songs.

Statistic	RWB	YLW	MWR	SWR	FOX	SON
Mean	0.89	1.05	1.58	1.71	2.49	2.93
Standard deviation	0.17	0.31	0.42	0.22	0.28	0.45
Number of songs	17	27	32	13	13	31

Times are in seconds.

2.4 Discussion

This study can be described as speaker-independent song recognition. Bird songs used in this study had been recorded from individuals in a variety of locations, with differing quality and amounts of background noise. The bimodal song length distributions of some species may indicate that these species have two classes of songs, or that there are two dialects in the data set. For the RWB songs, a human listener can detect distinct song classes. The overall performance achieved with the 512 sample window was good; increasing the window size or combining two resolutions of spectral data resulted in a slight improvement in the performance of the neural network. Both of these modifications add some temporal context in addition to that provided by the length variable. The type of information available to the network due to pre-processing is not unlike that extracted by the vertebrate auditory system.

It is well known that the cochlea performs a time to frequency transformation. Recent work with bats (Suga, 1990) and barn owls (Konishi, 1993) points to sophisticated pre-processing of spectral data in vertebrates. Both appear to have neurons that serve as delay lines and coincidence detectors. Bats measure the time difference between sonar emissions and echoes, and thus can assess the distance between themselves and an object. The barn owl accurately locates a prey item if it makes a sound. It does so by comparing the time of arrival and amplitude of sounds reaching its two ears. Carr (1993) indicates that such delay elements may be widespread in the nervous system of vertebrates. Dear *et al.* (1993) have shown that bats have neural mechanisms to assemble a complete sonar image 40 msec after emitting a pulse. They do so because the neurons activated by the rapidly returning echoes have a long latency; they remain activated until the late-arriving echoes arrive.

It is not unreasonable to expect that populations of neurons with latency variation could permit vertebrates to compile an integrated picture of a short sound. This suggests that vertebrates may also be able to combine short-term and long-term spectral information,

and easily learn to distinguish sounds. The neural network / pre-processing scheme used in this study is a simple model of such a system.

Although the methods described in this section provided good results, initial results did not look as promising. These methods were also computationally intensive; since many frames were used for training the neural networks, training took several hours on a high performance workstation. For these reasons, another means of pre-processing bird songs was investigated; this approach is described in the following section.

3. SPECIES IDENTIFICATION - METHOD TWO

3.1 Introduction

Literature and previous research (Ch. 1 and Ch. 2) indicate that bird sounds are variable and often complex. This can make the task of identifying species based on their sounds difficult. Species identification shares many of the difficulties encountered in the related tasks of human speech recognition and speaker identification. The foremost problem is that of variability. One finds that groups of individuals have dialects, that individuals are distinct, and that individuals exhibit considerable variation in the sounds they make. Another difficulty is the volume of data generated as a result of sampling sounds. Depending on the frequency, bandwidth and amplitude resolution desired, one may generate up to 88k bytes per second of recording (single channel, 44.1k samples/sec, monophonic, 16 bits per sample). In order to automatically identify birds by their sounds, it is thus necessary to devise methods which tolerate variability, reduce dimensionality and still provide reliable identification. It is a premise of this research that pre-processing sampled sounds to extract simple features is a prerequisite to reliable discrimination.

Certain temporal features of bird calls have been shown to be important to both birds and humans for identifying species. Some of these are relatively easy to extract automatically. One can assume, for example, that a bird song is made up of periods of sound (elements) and of silence. It should be possible to parse the songs by using an algorithm that detects the beginning and end of elements. The length and number of such features is known to be a useful clue about species identity. Once a song is parsed, one is at liberty to discard silent periods and to examine the spectral qualities of each element.

Again, it appears from the literature that coarse spectral features are useful in discriminating among species. Even with the removal of silence, one is left with a large number of samples to analyze. By calculating the power spectral density (PSD) for each element, it is possible to produce a small amount of coarse spectral information for each song. This information coupled with temporal information provides a basis for discriminating between bird species based on their sounds.

Once feature variables are generated, it is advisable to examine them in an exploratory manner. One can determine if variables (or a transformed version of them) are normally distributed, inter-correlated or heteroscedastic (having non-uniform variance among classes). Based on this information, a variety of appropriate statistical methods may be applied to further examine data structure and to choose a reduced number of variables for subsequent analyses. Lastly, statistical and neural network discrimination methods can be used to discriminate between bird species based upon the pre-processed data.

The purpose of this chapter is to provide an alternative means of analyzing bird sounds that requires less computation than that described in Ch. 2, and that is amenable to exploratory statistical analysis. A technique for automatically parsing bird sounds, and for generating a set of coarse feature variables is described first. Statistical methods are used to explore the data structure of these variables, and to choose a subset of them for the discrimination task. Finally, artificial neural networks and discriminant analysis are compared as methods for identifying unknown bird species based on their sounds.

3.2 Methods

3.2.1 Sampling

Sampled bird songs used in this study were the same as those used in the previous chapter. C++ programs were written to read binary wave files (windows *.wav format) containing the samples, and to perform the pre-processing described herein.

3.2.2 Pre-processing

Parsing of a bird sound was accomplished using a 'leaky-integrator' algorithm. Each sample of the sound was read, and if its absolute magnitude exceeded a value (threshold), the integrator was incremented by a constant (attack). If the magnitude fell below the threshold, the integrator was decremented by another constant (decay). The integrator was not allowed to exceed a constant (limit) in either the positive or negative direction. By adjusting the attack and decay parameters it was possible to parse songs in a manner that matched human perception of elements and silences within a song; as a result of adjustment, silences that were too short in duration to be perceptible were ignored. By adjusting the threshold, sensitivity to noise was reduced.

The parsing program determined the number of elements in a song, and calculated the mean and variance of both element and silence lengths within each song. In order to remove some amplitude variation prior to spectral analysis, the parsing program also normalized the amplitude of each element to the maximum scale for the data format (+127, -128 for this 8-bit data). To verify that this procedure did not adversely affect the song, normalized songs were played back; the only perceptible difference was an overall increase in volume. Five temporal variables were produced for each song: mean and standard deviation of element length, mean and standard deviation of silence length, and the number of elements.

In order to obtain a statistically robust estimate of a song's frequency content, the PSD was calculated for each element. The Welch method (Press et al., 1992) was applied using a triangular time-domain window and 16 point FFT's. Variation in element lengths forced the number of FFTs (M) to vary from element to element. The squared magnitudes for the nine spectral bands (where redundant negative frequencies were excluded) were accumulated. If the last window of an element contained less than 16 samples, data from this window was discarded. Once results from all FFTs were accumulated, the totals were

divided by M and their square root taken. The final band averages were then normalized with respect to the band containing the largest average; bands 1 - 9 covered the range of 0 - 5.125kHz. This was done to further reduce the effect of amplitude variations in the original signal. Means and standard deviations for each of the nine spectral bands constituted the eighteen spectral variables produced from each song.

3.2.3 Statistical analysis

The data set that resulted from pre-processing contained 23 variables in 133 records. This was uploaded to the UNIX network, where SAS (1985a,b) software was used for all statistical analyses.

Preliminary examination of the correlation structure of the data indicated complex inter-correlation of variables. This in turn suggested that a smaller set of variables was expected to contain enough information to permit discrimination. A stepwise discriminant analysis was performed using the STEPDISC procedure of SAS. The significance criterion used to enter a variable and to retain it during the elimination phase was $\alpha = 0.15$.

Given that all records belonged to one of six labeled groups, it was possible to perform a single factor ANOVA with 6 factor levels (the species classes) using each of the eight variables. For many model based statistical methods, including ANOVA, it is desirable that data have a normal distribution and a variance that is equal in each class (homoscedasticity) If such assumptions are not met, extra care must then be taken in interpreting results, especially if the goal is hypothesis testing. In this study, separate ANOVAs, and Scheffe's multiple comparison test (Neter and Wasserman, 1974) were calculated for each variable. Residuals were checked for normality using the Shapiro-Wilk statistic W calculated within the UNIVARIATE (SAS, 1985a), and homoscedasticity with Bartlett's Test (Neter and Wasserman, 1974) using a SAS macro (Friendly, 1995). It is known that application of a monotonic transformation to raw data may sometimes improve the distributional or variance properties of the data. The same procedures were applied to the log and square root of the raw data. In order to visually verify that ANOVA results were reasonable, Tukey box plots were generated for each variable using UNIVARIATE (SAS, 1985a).

Another way of exploring the structure of data is to use ordination methods. The amount of overlap among songs of the six bird species was examined by reducing the eight variable data set to a two-dimensional space using PCA and canonical discriminant analysis (CDA) with PRINCOMP and CANDISC of SAS. The correlation option was used for PCA, so that each variable was given equal weight in analysis regardless of its scale.

Finally, classification discriminant analysis (DISRIM of SAS) was applied to the reduced data set. Prior probabilities were assumed to be equal. Preliminary tests with DISCRIM indicated that quadratic discrimination functions were necessary since covariance matrices for different classes were too dissimilar to permit pooling. A SAS program was

written to divide the 133 record data set in stratified random manner into a 'test' set containing one record of each species, and a 'training' set which contained the remaining 127 records. Training data were used to generate discriminant functions, which were then used to classify the 'unknown' observations in the test set. In order to reduce the impact of atypical records, results were based on 25 different randomizations.

3.2.4 Neural network analysis

In an attempt to optimize the network configuration and training time, training and test sets were generated in the same manner as that used in Ch. 2. 'Vanilla' back-propagation, without momentum or higher order derivative information was used as the learning model, as it was in Ch. 2. The initial learning rate was set to 0.2, and target values of 0, 1 were changed to 0.1 and 0.9 to accelerate learning and prevent weight inflation; since these values are larger than those used in Ch.2, training time is increased slightly. Training was accomplished using the PDP (McLelland and Rumelhart, 1989) back propagation algorithm on a PC. The number of inputs and outputs was set at eight and six respectively. Network and training parameters were varied to find a configuration that would optimize generalization performance. To this end, the number of hidden nodes was varied from three to eight, and the number of training epochs from 20 to 500. Classification accuracy was determined using the independent test data. Every configuration was evaluated with 5 trials, each utilizing different randomly selected test sets, training sets and initial network weights.

A post-processing program was written to read in the weights generated from a training episode, and to compute the six output activations for each record in the test set. For a given record, any output with an activation > 0.6 was considered to be 'on'. If any incorrect output was on or if no outputs were on, a misidentification was tallied.

For actual identification trials, training and test sets were generated in the same manner as that used with classification by discriminant analysis above. This was done to make the results comparable to those analyzed by that method.

3.3 Results

3.3.1 Statistical analysis

Based on the results of stepwise discriminant analysis (Table 3.1), the number of variables used in subsequent analyses was reduced from 23 to eight. The average squared canonical correlation indicated that only small gains in discrimination (compare steps eight and nine) were likely if more than eight variables were retained. Using fewer variables eased further examination of the data.

Table 3.1 Stepwise discriminant analysis of temporal and spectral variables. SD is standard deviation, ASCC is average squared canonical correlation.

Step	Variable added	Variable removed	ASCC
1	Mean element length (XE)	-	0.14
2	SD of band 4 (S4)	-	0.24
3	Mean silence length (XS)	-	0.33
4	Number of elements (NE)	-	0.42
5	SD of element length (SE)	-	0.46
6	Mean for band 3 (X3)	-	0.49
7	Mean for band 2 (X2)	-	0.52
8	Mean for band 8 (X8)	-	0.56
9	SD of band 7 (S7)	-	0.57

The relative contribution of each of the chosen variables is indicated in Table 3.2. Those variables with large partial correlations and F-statistics contribute most to discrimination. Note that all F-statistics have a low probability of occurring due to chance ($p \leq 0.0001$). XE and NS are the dominant temporal variables; S4 is the dominant spectral variable. The eight variables chosen at this stage are used in all subsequent analyses.

Table 3.2 Contribution of variables to the discriminant model, including the partial correlation coefficient for the variable, and the corresponding F-test for significance.

Variable	Partial r^2	F	Prob > F
XE	0.52	26.2	0.0001
S4	0.59	35.0	0.0001
XS	0.42	17.1	0.0001
NE	0.57	32.2	0.0001
SE	0.26	8.5	0.0001
X3	0.39	15.1	0.0001
X2	0.39	15.4	0.0001
X8	0.26	8.4	0.0001

ANOVA results and tests of assumptions are summarized in Table 3.3. Hypothesis tests with such data tend be misleading (Neter and Wasserman, 1974) since the assumptions of

normality and homoscedasticity are violated in many cases. Although transformations were attempted, they generally improved performance on one assumption while worsening it for the other. Given that the goal at this stage was to simply explore similarities and differences among species, this was not considered to be a large problem so long as the general trends were not misleading. Inspection of Tukey box plots (which are included and explained in the appendices) confirmed that the pattern of differences among species suggested by the Scheffe's test results were reasonable.

Note that each variable yielded a different pattern of overlap among species, reinforcing the idea that the chosen variables each contribute different information that can be used to discriminate between species. Fox Sparrows and Marsh Wrens, for example, cannot be distinguished by the number of elements in their calls but they could be with mean element length.

Table 3.3 Analysis of variance summary. Species with same letter in a column have means that do not differ significantly (Scheffe's test, $\alpha = 0.05$) .

Species	XE	S4	XS	NE	SE	X2	X3	X8
FOX	A	A	A	A	A	A	A	A
MWR	B	A	A	A B	B	A	A B	A
RWB	B C	A B	A	C B	B	A B	C B	B
SON	B C	C B	B	C D	B	A B	C	B C
SWR	B C	C	B C	D	B	B	C	B C
YLW	C	C	C	E	B	B	C	C
Normal*	no	yes	no	no	no	no	no	yes
Homoscedastic*	no	no	no	no	no	no	no	yes

* Tests are at the $\alpha = 0.10$ level of probability.

The results of a principal components analysis of the eight variables are given in Tables 3.4, 3.5 and Fig 3.1. Standardized variance accounted for by the first three components is 69%, indicating that the data has a strong underlying pattern. This also indicates that there is still some redundancy in the information provided by the original eight variables.

Table 3.4 PCA, using correlation, for the eight variable data set. The proportion of standardized variance accounted for by each axis and the cumulative total are indicated.

Axis	Eigenvalue	Proportion	Cumulative
PC1	2.64	0.33	0.33
PC2	1.73	0.22	0.55
PC3	1.18	0.15	0.69
PC4	0.83	0.10	0.80
PC5	0.39	0.09	0.88
PC6	0.50	0.06	0.95
PC7	0.28	0.04	0.98
PC8	0.15	0.02	1.00

The eigenvectors resulting from PCA provide additional useful information about the relative contribution of each variable to a given PCA axis. Based on the two vector components with the highest absolute magnitude for each axis (Table 3.5), PC1 appears to be dominated by spectral variables, PC2 by temporal ones, and PC3 by both; this pattern is not particularly strong, however.

Table 3.5 Eigenvectors for the first three PCA axes; the two largest eigenvector components are indicated in bold type.

Variable	PC1	PC2	PC3
XS	-.23	0.20	**0.72**
XE	0.34	**-.50**	0.06
SE	0.37	-.27	0.13
NE	-.15	**0.62**	-.18
X2	0.36	0.23	**0.52**
X3	**0.45**	0.23	-.06
S4	0.37	0.28	-.37
X8	**-.45**	-.27	-.09

Although there is considerable overlap among the six species groups, Fig. 3.1 indicates that Red-winged Blackbirds are best separated along PC1 and PC2. The RWB group is marginally less distinct along PC3. Fox Sparrows, however, are pulled away from Marsh

Wrens along PC3. Yellow Warblers, Sedge Wrens and Song Sparrows are best separated when trends along all three axes are considered together.

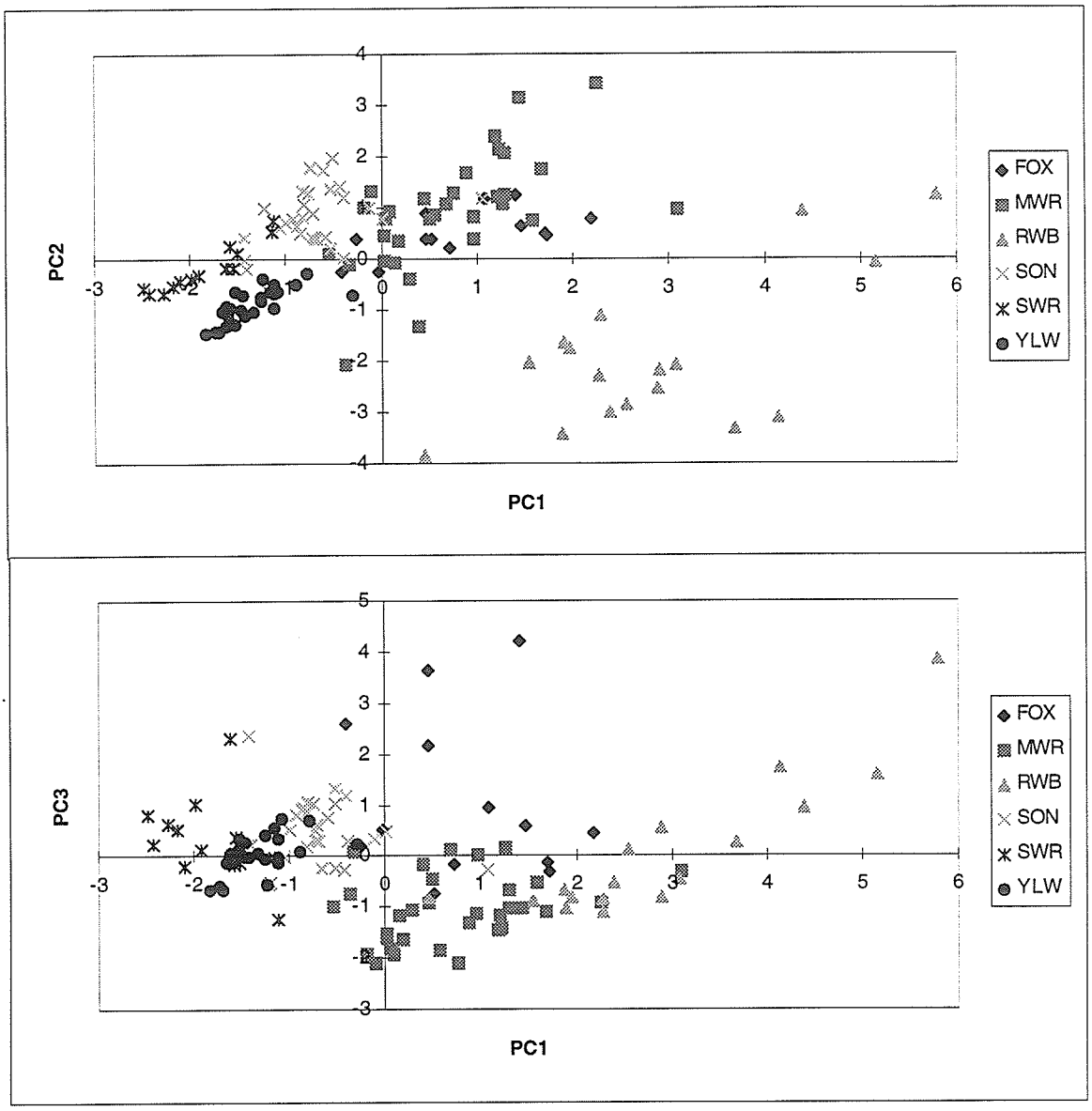


Fig. 3.1 The first three principal components from the 8 variable data set, plotted by species.

Canonical discriminant analysis yielded a much stronger data structure than PCA, as was expected (Tables 3.6 and 3.7). Given information about the identity of the classes, CDA was able to account for 83% of standardized variance with only two axes, and all of it with five. All variables contributed to CD1 , with the time domain variables element length and standard deviation being the most prominent. CD2, like PC3 is affected by both temporal and spectral variables.

Table 3.6 CDA for the eight variable data set. The proportion of standardized variance accounted for by each axis and the cumulative total are indicated.

Axis	Eigenvalue	Proportion	Cumulative
CD1	10.19	0.68	0.68
CD2	2.38	0.16	0.83
CD3	1.7	0.11	0.95
CD4	0.62	0.04	0.99
CD5	0.22	0.01	1.00

Table 3.7 Eigenvectors for the two CDA axes; the two largest eigenvector components are indicated in bold type.

Variable	CD1	CD2
XS	-.63	0.56
XE	**0.94**	-.007
SE	**0.92**	0.12
NE	-.64	**0.58**
X2	0.73	0.48
X3	0.69	0.003
S4	0.60	0.41
X8	-.66	**-.61**

Again as expected, CDA results in a clearer separation of species groups (Fig. 3.2), with Yellow Warblers and Red-winged Blackbirds forming quite distinct groups. The other four species show varying degrees of overlap, with Fox Sparrows and Marsh Wrens being the least distinct.

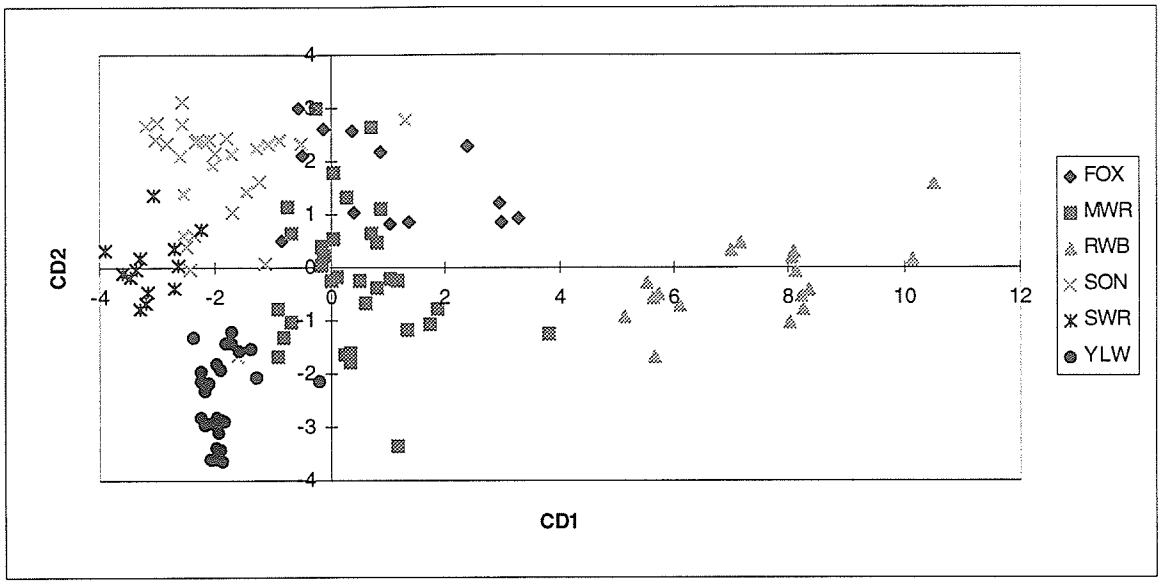


Fig. 3.2 The first two canonical discriminants from the 8 variable data set, plotted by species.

Since the covariance matrices of each species group were too dissimilar to pool, quadratic discriminant functions were used in classification discriminant analysis. The results (Fig. 3.3) for the *test* records were excellent, and showed a similar pattern of variation to those observed in Ch.2. Overall accuracy was 93.3%. No errors were made in assigning *training* records to their proper categories.

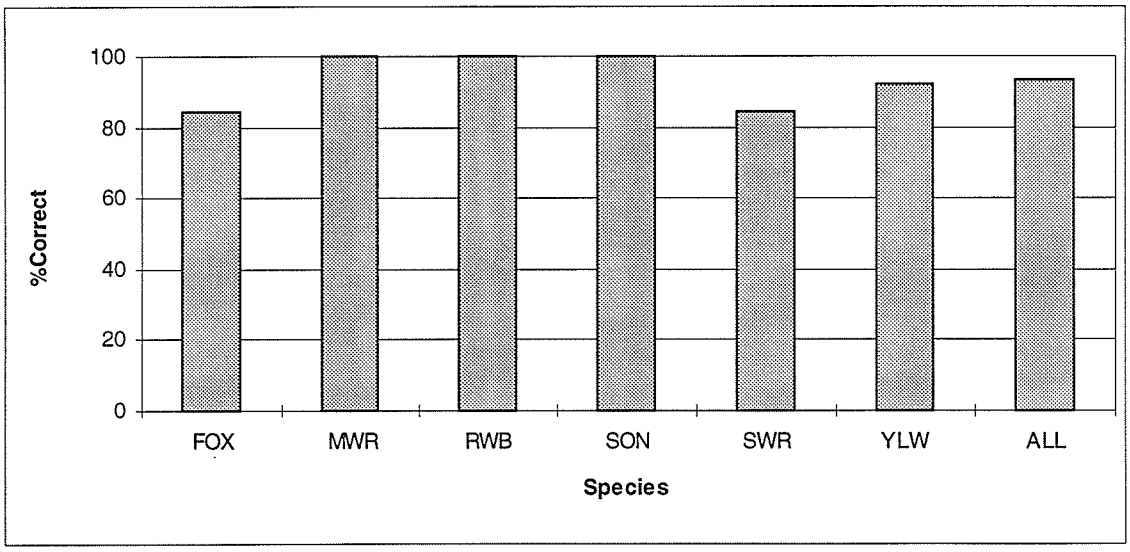


Fig. 3.3 Identification accuracy with discriminant analysis for the six species and the total.

3.3.2 Neural network analysis

In order to find a configuration for the neural network that was likely to yield performance that was close to the optimum, a range of network configurations and

training times was explored (Fig. 3.4) using the same eight variable data as before. Since preliminary tests indicated that training for 200 epochs resulted in reasonable generalization, networks with 3-8 hidden units were evaluated. Accuracy fluctuated little; since training times were in the order of minutes on a 386DX-33Mhz PC, the configuration that gave marginally the best accuracy (6 hidden units) was chosen for further tests.

The number of epochs required to achieve the best generalization on test data rose consistently until the 80 epoch period was reached. After that fluctuations were probably insignificant. Due again to the fact that training times were short, 200 epochs (the point of maximum accuracy) was chosen as the training period to be used in further analyses.

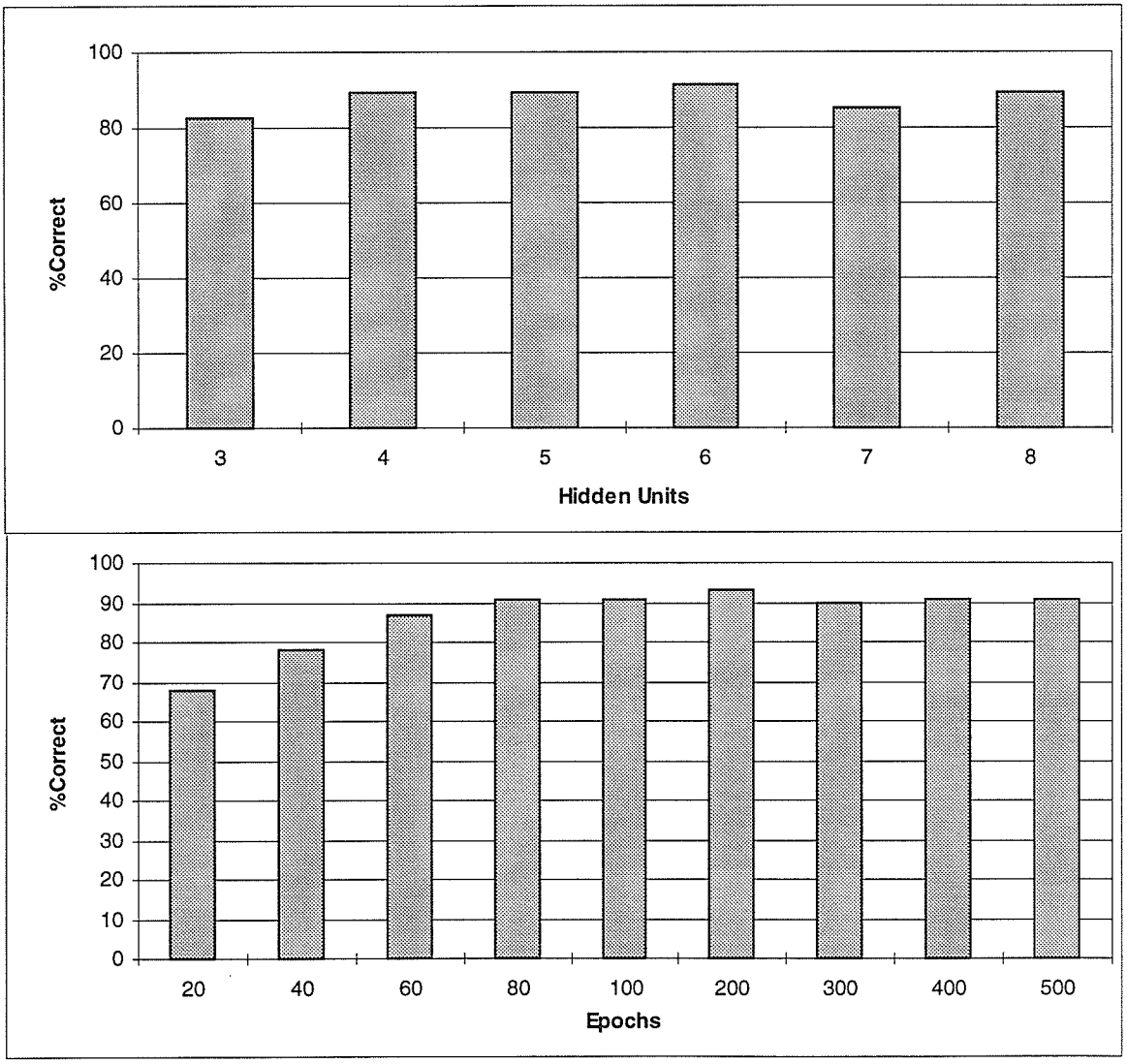


Fig. 3.4 Effect of number of hidden units, and length of training period on generalization of back-propagation neural networks. Six hidden units were used in evaluation of training periods.

Running the chosen network configuration (eight inputs, six hidden units and six outputs for 200 epochs) repeatedly, with one record of each species used for testing, yielded the results indicated in Fig. 3.5. These results are not as good as those indicated for discriminant analysis but do show similarities in trends by species; Red-winged Blackbirds show the best performance, and Sedge Wrens the poorest. Overall accuracy was 82%.

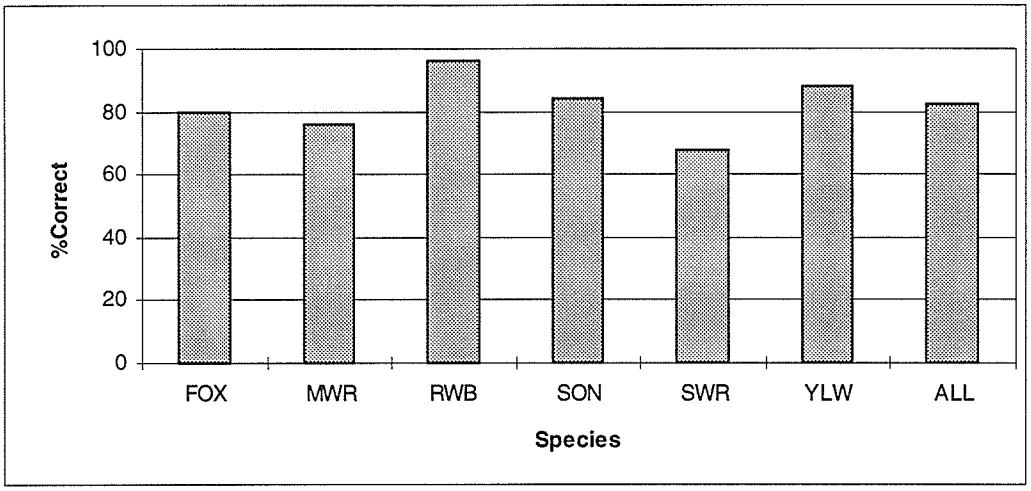


Fig. 3.5 Identification accuracy with a back-propagation neural network for the six species and the total.

3.4 Discussion

Fig. 3.6 provides a summary of the process used in this chapter for species recognition, and emphasizes the complementary nature of statistical and neural network methods.

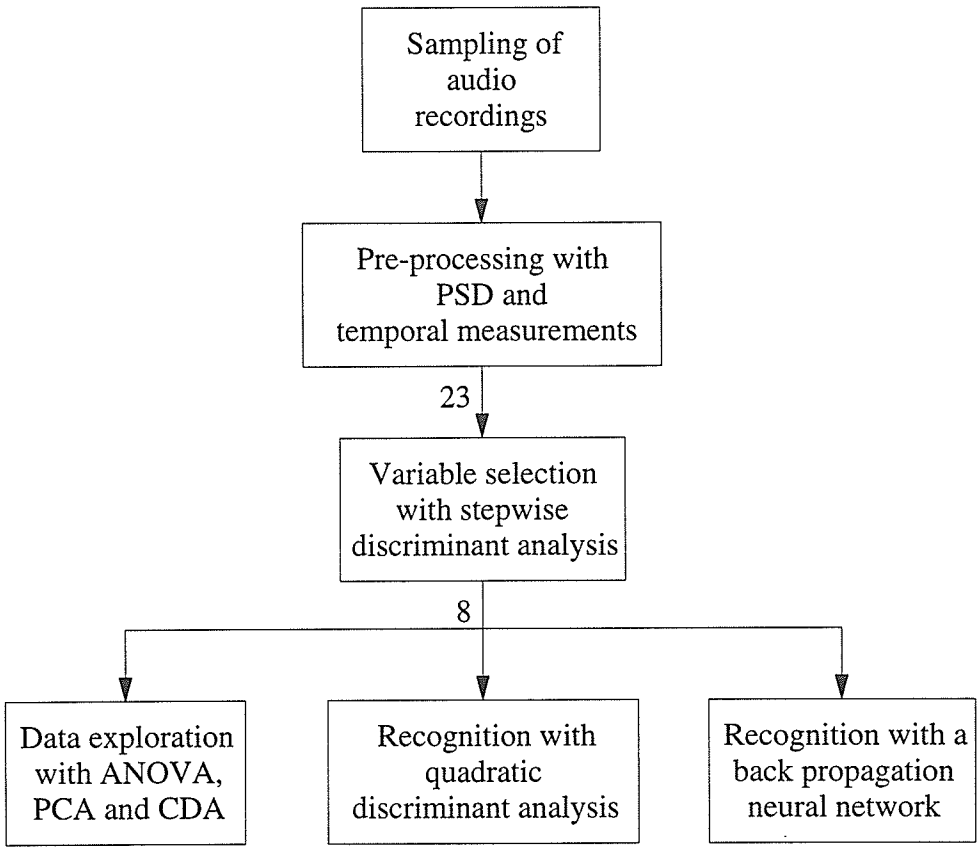


Fig. 3.6 Overview of analysis procedure. Numbers next to lines indicate the number of variables at each step.

3.4.1 Exploration of the data

As suggested by James and McCulloch (1990) care must be taken in speculation about the theoretical significance of the variables chosen by stepwise procedures. In this study, such methods were used to choose a subset of the original 23 variables since there was much inter-correlation among the original ones. Had all variables been retained, computational loads would have been higher in subsequent analyses, and performance would have been lower due to the large number of parameters that would have been involved in the models. It is not surprising that both temporal and spectral variables were discovered to be good predictors of species identity. As discussed in Ch.2, animals use and require both temporal and spectral information to accomplish their many tasks. In fact these cues are also believed to be important for species recognition among birds (Becker, 1982).

Together, the four temporal variables selected provide information on the number of elements, their mean length and their variability, and the mean length of silent periods. The fact that this almost exhausts the list of measured temporal variables, and that their associated F-statistics were large suggests that these six species can be distinguished largely by their patterns of sound versus silence. Spectral measurements are required to sort out fine distinctions. Using a larger number of species would make the recognition task more difficult, and could also alter the relative contribution of the variables. It is possible that a hierarchical network of experts (Jacobs et al., 1991) may have to be employed to identify different species groups using different variables sets as alluded to in section 1.3.2.

Given the eight variables that were chosen, one then becomes interested in what information each variable contributes to the task of discrimination. Analysis of variance suggests that each variable on its own could be used to discriminate certain subgroups within the six species. Each variable seems to contribute different information, as it should if stepwise variable selection was used effectively to choose a relatively uncorrelated subset of the original variables.

ANOVA model assumption violations resulted from a number of zero standard deviations generated for RWB songs, where only one element was detected. It is also believed that some species contained song subgroups; this would create bimodal or other distributions that violated the assumption of normality. Non-normality is not a serious problem as long as data are homoscedastic and sample sizes for each class are similar (Neter and Wasserman, 1974). Unfortunately this was not always the case. Although the un-transformed data violated ANOVA model assumptions, the patterns were not unreasonable when compared with Tukey box plots of the same data. Formal hypothesis tests were not performed.

ANOVA, in the form used here, provided some insight into the characteristics of the eight variables separately. PCA and CDA provided a graphical two- or three-dimensional representation from the multivariate data. Eigenvalues for both analyses suggest that there was a strong underlying structure in the data set. In other words, it is unlikely that the patterns observed arose by chance. Eigenvalues also suggested that linear transformation of the eight variables into two or three variables retained much of the information. Eigenvectors showed that trends along each axis were dominated by different variables, and that the two techniques differed in having temporal variables dominating the first axis (CDA) or spectral variables in the case of PCA. In CDA, temporal variables heavily influenced axis 1, suggesting again that they are strong contributors to discrimination. The plots for PCA and CDA indicate that there is much overlap of the species in a multivariate sense, with one species (RWB) being clearly separated from the others; this separation is not visible in the univariate ANOVA results. The ability of CDA to find linear combination of variables that emphasized the differences (SAS, 1985b) between species, permitted fewer axes to be used and increased the degree of separation between species.

All of this information suggests that the pre-processing methods were successful in extracting relevant information for discrimination of bird species. The eight variable data, however, contains considerable multivariate overlap between species, indicating that the identification task was tractable but not trivial.

3.4.2 **Identification**

Classification discriminant analysis identifies records based on characteristics of covariance matrices and works best if data are multivariate normal (SAS, 1985b). This means that the distributional problems that affected ANOVA results may have had a detrimental effect on discrimination. This effect is softened however because quadratic discrimination uses a separate covariance matrix for each species rather than a pooled estimate. The excellent results, however, indicate that the effect of data distribution on performance was small.

When the nature of the errors that discriminant analysis made in classifying was examined, certain records (five of them) were found to be responsible for the errors. Examination of these suggested that they did indeed sound different from the others, but that it was difficult to discern what the critical differences were. These pathological samples were not incorrectly labeled, but could have represented atypical or incomplete songs. They also may represent the presence of uncommon dialects among the sample songs used in this study. Such results again emphasize the need for large and representative samples in this type of research.

The neural network classifier did not perform as well as quadratic discriminant analysis, or as well as the network described in Ch.2. The pattern among species is consistent with other results in this paper, but the overall performance was lower. Examination of trends in overall accuracy with respect to the number of training runs indicated that performance declined between 15 and 25 runs. It is possible that pathological cases were more prevalent among samples. Further tuning and a larger number of runs could, in theory, result in network performance that is similar to that for discriminant analysis. Even so, the low number of hidden units and short training periods that were required reinforce the idea that the data contained clear structure. It also appears that the pre-processing methods resulted in the requirement for a much simpler network than that of Ch.2.

In comparison with the work done in the previous chapter, these methods use explicit temporal preprocessing and statistical methods to greatly reduce the volume of data presented to the final discriminator. In addition they permit detailed exploration of the data. Lastly, since quadratic discriminant analysis is a simple method that requires no tuning, species identification becomes simpler. For these reasons, I feel that methods in this chapter provide a better means for identifying birds by their songs.

Although this research has demonstrated that six bird species could be identified automatically by their songs, it poses a number of further questions and problems:

a) How would performance decline as more species were added to the data set?

b) What is a reasonable sample size for a given number of species?

c) What effect do dialects and other variations have on performance?

d) Could unsupervised methods detect such groups?

e) Could further tuning and testing improve the performance of the neural network?

f) Could tuning of pre-processing methods improve performance and distributional characteristics of variables?

g) Does the choice of variables change as other species or more samples are added?

h) What happens when sounds other than those of the six species are presented to the identifier?

i) Are there better methods for this task?

As a note in passing, I know of only one other published description of an artificial neural system designed to recognize bird calls. Ashiya and Nakagawa (1993) used back-propagation on spectral data; although the description of their method was somewhat sketchy, they appeared to obtain 87.5% accuracy using nine species.

4. Conclusion

A number of procedures were successfully applied to the task of automatically recognizing bird species by their sounds. The first approach focused on spectral characteristics (LPC coefficients) of sub-sections of bird songs and identified the song by counting the number of frames that were characteristic of each species. A neural network successfully identified the bird songs using this technique.

The second approach involved extracting temporal and spectral (PSD) features, using univariate and multivariate statistics in variable selection, and using discriminant analysis and neural networks for final identification. The computational time required was considerably reduced when compared with the first approach. Again the results varied from being good to excellent.

An inter-disciplinary approach drawing information from biology, computer engineering and statistics was used in this research. Together, these fields provided both the tools and the clues that made it possible to design these automatic recognition systems. In fact, classic statistical techniques provided the means to critically evaluate the utility of the variables that were used for recognition. There is considerable research still remaining to be done on this topic, which could lead to the development of practical implementations in the areas of biological research and education.

5. LITERATURE CITED

Amari, S. 1994. "Comment", added to "Neural networks: a review from a statistical perspective", by Cheng, B. and D.M. Titterington, *Statist. Sci.* 9(1)2-54.

Anderer, P., B. Saletu, B. Kloppel, H.V. Semlitsch and H. Werner. 1994, "Discrimination between demented patients and normals based on topographic EEG slow wave activity: comparison between *z*-statistics, discriminant analysis and artificial neural network classifiers," *Electroenceph. Clin. Neurophysiol.* 91:108-117.

Ashiya, T. and M. Nakagawa. 1993, "A proposal for a recognition system for the species of birds receiving birdcalls - an application of recognition systems for environmental sound," *IEICE Trans. Fundamentals* E76-A:1858-1860.

Atal, B.S. 1985, "Linear predictive coding of speech", in *Computer Speech Processing*, F. Fallside and W.A. Woods, eds. Prentice-Hall, Inc., London, pp. 81-124.

Aubin, T. 1992, "Some features of time-frequency analysis and representation of animal vocalizations", in "The conference report of the XII Symposium of the International Bioacoustics Council," *Bioacoustics* 4:59-60.

Baker, M.C. 1983, "Sharing of vocal signals among Song Sparrows," *Condor* 85:482-490.

Bankwitz, K.G. and W.L. Thompson. 1979. "Song characteristics of the Yellow Warbler," *Wilson Bull.* 91:533-550.

Becker, P.H.1982, "The coding of species-specific characteristics in bird sounds," in *Acoustic Communications in Birds, Vol. I*, D.E. Kroodsma, E.H. Miller and K. Ouellet, eds. Academic Press, N.Y. pp.213-252.

Benediktsson, J.A., P.H. Swain and O.K. Ersoy. 1990, "Neural network approaches verses statistical methods in clasification of multisource remote sensing data," *IEEE Trans. Geosci. Remote Sens.* 28(4):540-552.

Bhattacharyya, G.K. and R.A. Johnson. 1977, *Statistical Concepts and Methods*. Wiley, New York. 639p.

Borror, D.J. 1967, *Common Bird Songs.* Dover, New York.

Borror, D.J. 1970, *Songs of Eastern Birds*. Dover, New York.

Brackenbury, J.H. 1982, "The structural basis of voice production and its relationship to sound characteristics," in *Acoustic Communications in Birds, Vol. I*, D.E. Kroodsma, E.H. Miller and K. Ouellet, eds. Academic Press, N.Y. pp. 53-73.

Brackenbury, J.H. 1989, "Functions of the syrinx and the control of sound production," in *Form and Function in Birds, Vol. 4*, A.S. King and J. McLelland, eds. Academic Press, N.Y. pp. 193-220.

Breiman, L. 1994, "Comment," added to "Neural networks: a review from a statistical perspective" by Cheng, B. and D.M. Titterington, *Statist. Sci.* 9(1)2-54.

Brigham, M. No date, *Bird Sounds of Canada*. Holborne Dist. Co. Ltd, Mount Albert, Ontario.

Bryan, K., R. Moldenhauer and D.E. Kroodsma. 1987, "Geographic variation in songs of the Prothonotary Warbler," *Wilson Bull.* 99:396-376.

Byrne, J.H. 1987, "Cellular analysis of associative learning," *Physiol. Rev.* 67(2):329-439.

Canadian Wildlife Service. 1980, *Migratory Birds Convention Act*, R.S., 1970, c. M-12 and the migratory birds regulations established by C.R.C., c. 1035 and amendments. Canadian Wildlife Service, Department of the Environment, Ottawa, Ontario, Published by Minister of Supply and Services Canada.

Card, H.C. and W.R. Moore. 1990, "Silicon models of associative learning in *Aplysia*," *Neural Networks*, 3:333-346.

Carr, C.E. 1993, "Processing of temporal information in the brain," *Annu. Rev. Neurosci.* 16: 223-263.

Cheng, B. and D.M. Titterington. 1994, "Neural networks: A review from a statistical perspective," *Statist. Sci.* 9(1)2-54.

Curtis, H. 1983, *Biology*, 4th ed. Worth Publishers, Inc., N.Y. 1159p.

Date, E.M., R.E. Lemon, D.M. Weary and A.K. Richter. 1991, "Species identity by birdsong: discrete or additive information?," *Anim. Behav.* 41:111-120.

Dear, S.P., J.A. Simmons and J. Fritz. 1993, "A possible basis for representation of acoustic scenes in auditory cortex of the big brown bat," *Nature* 364:620-623.

Denes, P.B. and P.N. Elliot. 1993, *The Speech Chain*, 2nd ed. W. H. Freeman and Co, New York.

Deutsch, D.1992, "Paradoxes of musical pitch," *Sci. Am.* 267:88-95.

Dooling, R.J. 1982, "Auditory perception in birds," in *Acoustic Communications in Birds, Vol. I*, D.E. Kroodsma, E.H. Miller and K. Ouellet, eds. Academic Press, N.Y. pp. 95-129.

Dror, I.E., M. Zagaeski and C.F. Moss. 1995, "Three-dimensional target recognition via sonar: a neural network model," *Neural Networks* 8:149-160.

Duda, R.O. and P.E. Hart. 1973, *Pattern Classification and Scene Analysis*. Wiley, N.Y. 482p.

Eckert, R. and D. Randall. 1983, *Animal Physiology*, 2nd ed. W.H. Freeman and Co., San Francisco. 830pp.

Elliot, L. and T. Mack. 1990, *Wild Sounds of the Northwoods*. NatureSound Studio, Ithaca, N.Y.

Endler, J.A. 1992, "Signals, signal conditions and the direction of evolution," *Am. Nat.* 139(Supplement):S125-S153.

Evans, M.R. and J.A. Evans.1994, "A computer-based technique for the quantitative analysis of animal sounds," *Bioacoustics* 5:281-290.

Everitt, B.S. 1993, *Cluster Analysis*, 3rd ed. Edward Arnold, London. 170p.

Farabaugh, S.M. 1982, "The ecological significance of duetting," in *Acoustic Communications in Birds, Vol. II*, D.E. Kroodsma, E.H. Miller and K. Ouellet, eds. Academic Press, N.Y. pp.85-124.

Fletcher, N.H. 1988, "Bird song - a quantitative acoustical model," *J. Theor. Biol.* 135:455-481.

Friendly, M. 1995, "Bartlett's Test for Homogeneity of Variances: a SAS Macro" from *http://www.math.yorku.ca/SCS/friendly.html*. Department of Psychology, York Yniversity, Toronto, Canada.

Gaunt, A.S., S.L.L. Gaunt and T.M. Casey. 1982, "Syringeal mechanics reassessed: evidence from *Streptopelia*," *Auk* 99:474-494.

Godfrey, E.W.1986, *Birds of Canada*, Revised Edition. National Museum of Natural Sciences, Published by the National Museums of Canada, Ottawa. 595p.

Gorman, R.P. and T.J. Sejnowski. 1988, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks* 1:75-89.

Halliday, D. and R. Resnick. 1981, *Fundamentals of Physics*, 3rd ed. John Wiley and Sons, N.Y. 947p.

Handford, P. and S.C. Lougheed. 1991, "Variation in duration and frequency characters in the song of the Rufous-collared Sparrow, *Zonotrichia capensis*, with respect to habitat, trill dialects and body size," *Condor* 93:644-658.

Hanes, M.D., S.C. Ahalt, and A.K. Krishnamurthy. 1994, "Acoustic-to-phonetic mapping using recurrent neural networks," IEEE Trans. *Neural Networks* 5: 659-662.

Harris, R.J. 1975, *A Primer of Multivariate Statistics*. Academic Press, N.Y. 332p.

Haykin, S. 1994, *Neural Networks: a Comprehensive Foundation*. Macmillan College Publishing Co., N.Y. 696p.

Heil, P., G. Langner and H. Scheich. 1992, "Processing of frequency-modulated stimuli in the chick auditory cortex analogue: evidence for topographic representations and possible mechanisms of rate and directional sensitivity." *J. Comp. Physiol. A* 171:583-600.

Hertz, J., A. Krogh and R.G. Palmer. 1991, *Introduction to the Theory of Neural Computation*. Addison-Wesley Publ. Co., Redwood City, CA. 327p.

Hlawatsch, F. and G.F. Boudreaux-Bartels. 1992, "Linear and quadratic time-freqeuncy signal representations," *IEEE Sig. Proc. Magazine* 9:21-67.

Ingle, V.K. and J.G. Proakis. 1991, *Digital Signal Processing Laboratory Using the ADSP-2100 Microprocessor*. Prentice-Hall, Inc., London, 300p.

Jacobs, R.A., M.I. Jordan, S.J. Nowlan and G.E. Hinton. 1991, "Adaptive mixtures of local experts", *Neural Computation* 3:79-87.

Kellogg, P.P., R.T. Peterson and W.W.H. Gunn. 1975, *A Field Guide to Western Bird Songs*. Houghton - Mifflin, Boston.

Konishi, M. 1993, "Listening with two ears," *Sci. Amer.* 268:66-73.

Kroodsma, D.E. and J. Verner. 1978, "Complex singing behaviors among *Cisthothorus* wrens," *Auk* 95:703-716.

Kroodsma, D.E. and J. Verner. 1987, "Use of song repertoires among Marsh Wren populations," *Auk* 104:63-72.

Lemon, R.E., J. Struger and M.J. Lechowicz.1983, "Song features as species discriminants in american warblers (Parulidae) ," *Condor* 85:308-322.

Liebermann, P. 1984, *The Biology and Evolution of Language*. Harvard University Press, Cambridge, Mass. 379p.

Liebermann, P. and S.E. Blumstein. 1988, *Speech Physiology, Speech Perception and Acoustic Phonetics*. Cambridge University Press, Cambridge, UK. 249p.

Lippmann, R.P. 1989a, "Pattern classification using neural networks," *IEEE Comm. Magazine* 27(11): 47-64.

Lippmann, R.P. 1989b, "Review of neural networks for speech recognition," *Neural Computation* 1:1-38.

Manly, B.F.J. 1994, *Multivariate Statistical Methods: a Primer*, 2nd ed. Chapman & Hall, London. 215p.

Mardia, K.V., J. T. Kent and J.M. Bibby 1979, *Multivariate Analysis*. Academic Press, London. 521p.

Martin, D.J. 1977, "Songs of the Fox Sparrow. I. Structure of song and its comparison with song in other Emberizidae," *Condor* 79:209-221.

Martin, D.J. 1979, "Songs of the Fox Sparrow. II. Intra- and interpopulation variation," *Condor* 81:173-184.

Martin, D.J. 1990, "Songs of the Fox Sparrow III. Ordering os song," *Wilson Bull.* 102:655-671.

Martindale, S. 1980, "A numerical approach to the analysis of Solitary Vireo songs," *Condor* 82:199-211.

McGregor, P.K. 1991, "LSI's speech workstation: a sound analysis package for IBM PCs," *Bioacoustics* 3:223-234.

McKenna, G. 1995, "Ecological survey at Ontario Hydro's Nanticoke generating station: a biodiversity conservation initiative," *22nd Annual Edison Electric Institute Biologist's Task Force Workshop*. Harvey Hotel - Dallas, TX. Sept. 20-22.

McLelland, J.L. and D.E. Rumelhart. 1989, *Explorations In Parallel Distributed Processing*. MIT Press, Cambridge, Mass.

Mittelstaedt, M. 1994, "Ontario Hydro jolting employees into more productivity," *The (Toronto) Globe and Mail*, January 10, pp. A4.

Nelson, D.A. 1989, "The importance of invariant and distinctive features in species recognition of bird song," *Condor* 91:120-130.

Neter, J. and W. Wasserman. 1974, *Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Designs*. Richard D. Irwin, Inc., Homewood, Illinois. 842p.

O'Shaughnessy D. 1989, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Communications Magazine* 27:26-52.

Okanoya, K. and T. Kimura. 1993, "A software bird call detector and its applicaion to automated playback experiments," *Bioacoustics* 5:117-122.

Oppenheim, A.V. and R.W. Schafer. 1975, *Digital Signal Processing*. Prentice-Hall, Inc., Inc. Engelwood Cliffs, N.J. 585p.

Oppenheim, A.V., A.S. Willsky, and I.T. Young. 1983, *Signals and Sytems*. Prentice-Hall, Inc., London, 796p.

Papoulis, A. 1980, *Circuits and Systems*. Holt, Rinehart and Winston, Inc. Fort Worth, TX. 435p.

Payne, R.B. and P. Budde. 1979, "Song differences and map distances in a population of Acadian Flycatchers," *Wilson Bull.* 91:29-41.

Pettingill, Jr., O.S. 1985, *Ornithology in the Laboratory and Field*. 5th ed. Academic Press, Inc., Toronto, Ontario. 403p.

Pielou, E.C. 1984, *The Interpretation of Ecological Data: a Primer on Classification and Ordination*. John Wiley & Sons, N.Y. 263p.

Poon, P.W.F. and J.F. Brugge. 1993, "Virtual-space receptive fields of single auditory nerve fibers," *J. Neurophysiol.* 70:667-676.

Press, W.H., W.A. Teukolsky, W.T. Vetterling, and B.P. Flannery. 1992, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, N.Y. 994p.

Rabiner, L.R. 1994, "Applications of voice processing to telecommunications," *Proc. IEEE* 82:199-228.

Ralph, C.J., G.R. Geupel, P. Pyle, T.E. Martin and D.F. DeSante. 1993, *Handbook Of Field Methods For Monitoring Landbirds*. Gen. Tech. Rep. PSW-GTR-144. Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, Albany, California. 41p.

Rich, T. 1981, "Microgeographic variation in the song of the Sage Sparrow," *Condor* 83:113-119.

Ripley, B.D. 1994, "Neural networks and related methods for classification," *J. R. Statist. Soc. Series B.* 56(3):409-456.

Robbins, M.B., M.J. Braun and E.A. Tobey. 1986, "Morphological and vocal variation across a contact zone between the Chickadees *Parus Atricapillus and P. carolinensis*," *Auk* 103:655-666

Rumelhart, D.E., F.E. Hinton and R.J. Williams. 1986, "Learning representations by back-propagation of errors," *Nature* 323: 533-536.

Sachs, M.B. and N.A. O'Connell. 1983, "Frequency analysis in the central nervous systems of non-mammalian vertebrates," in *Bioacoustics : a Comparative Approach*, B. Lewis, ed. Academic Press, N.Y. pp. 283-310.

Sarle, W.S. 1994, "Neural networks and statistical models," in *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. SAS Institute, Cary, NC. 13pp.

SAS Institute Inc. 1985a, *SAS User's Guide: Basics*, Version 5 edition. Cary, NC 1290p.

SAS Institute Inc. 1985b, *SAS User's Guide: Statistics*, Version 5 edition. Cary, NC 956p.

Sataloff, R.T. 1992, "The human voice," *Sci. Amer.* 267:108-115.

Smith, D.G., F.A. Reid and C.B. Breen. 1980, "Stereotypy of some parameters of Red-winged Blackbird song," *Condor* 82:259-266.

Sparling, D.W. 1978, "Multivariate analysis of avian vocalizations," *J. Theor. Biol.* 74:83-107.

Spector, D.A. 1994, "Definition in biology: the case of "bird song" ," *J. Theor. Biol.* 168:373-381.

Srivastava, M.S. and E.M. Carter. 1983, *An Introduction to Applied Multivariate Statistics*. North Holland, New York. 394p.

Suga, N. 1990, "Biosonar and neural computation in bats,". *Sci. Amer.* 262:60-68.

Thielcke, G. 1969, "Geographic variation in bird vocalizations," in *Bird Vocalizations*, R.H. Hinde, ed. pp. 311-339. Cambridge University Press, Cambridge, UK.

Thompson, N.S., K. LeDoux and K. Moody. 1994, "A system for describing bird song units," *Bioacoustics* 5:267-279.

Thorpe, W.H. 1961, *Bird-Song: The Biology of Vocal Communication and Expression in Birds*. Cambridge University Press, London. 143p.

Tukey, J.W. 1977, *Exploratory Data Analysis*. Addison-Wesley, Reading , Mass.

Waibel, A. 1989, "Modular construction of time-delay neural networks for speech recognition," *Neural Computation* 1:39-46.

Walton, R.K. and R.W. Lawson. 1989, *Birding by Ear (Eastern / Central) - a Guide to Bird-Song Identification*. Houghton - Mifflin, Boston.

Weary, D.M., R.E. Lemon and S. Perreault. 1994, "Male Yellow Warblers vary the use of song types depending on pairing status and distance from the nest," *Auk* 111:727-729.

Weary, D.M., R.G. Weisman, R.E. Lemon, T. Chin and J. Mongrain. 1991, "Use of relative frequency of notes by Veeries in song recognition and production," *Auk* 977-981.

White, J.G., E. Southgate, J.N. Thomson and S. Brenner. 1986, "The structure of the nervous system of the nematode *Caenorhabditis elegans*," *Phil. Trans. R. Soc. Lond. B* 314:1-340.

Whitfield, I.C. 1994, *Neurocommunications: an Introduction*. John Wiley & Sons, N.Y. 304p.

6. APPENDICES

6.1 Envelope determination from LPC with Mathcad

LPC analysis of a 256 sample speech frame with Hypersignal

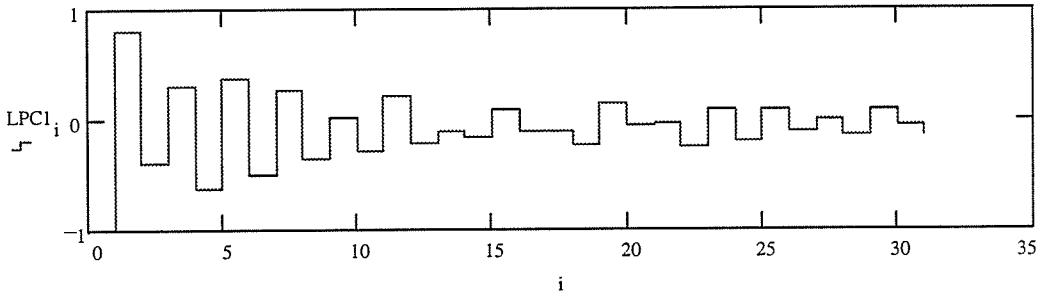
$i := 0..31$ $\quad LPC := \frac{READPRN(lpc_txt)}{1024}$

Read time domain coefficients exported from HSP and divide by 1024 to scale them to [-1,1]

$LPC_i := if(i<35, LPC_i, 0)$

$LPC1 := -LPC$

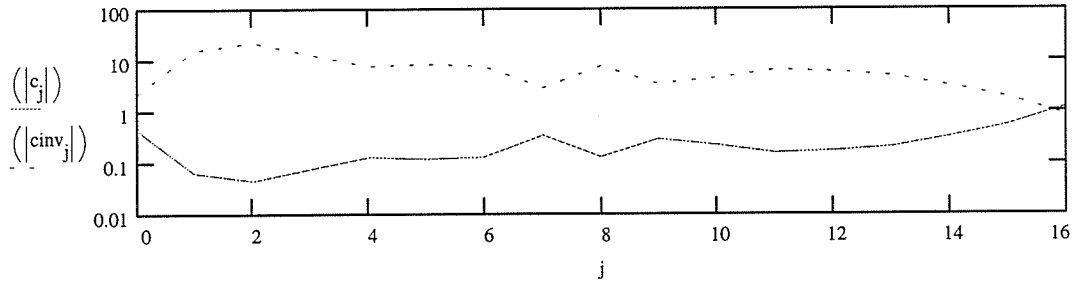
$LPC_0 := 1$



FFT of HSP's coefficients (this was used in bird song analysis), and inverse of FFT

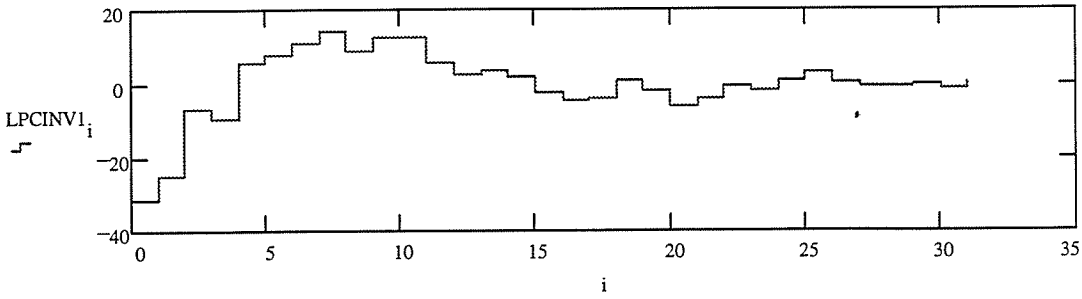
Real FFT $\quad c := fft(LPC1) \quad N := last(c) \quad N = 16 \quad j := 0..N \quad cinv_j := \frac{1}{c_j}$

If the LPC coefficients are treated as coefficients of an FIR filter, we can perform an FFT on them to determine the filter transfer function. This however provides a picture of the 'whitening' LPC filter (solid line) P(z). The estimated transfer function of the laryngeal filter (speech envelope) H(z), is actually obtained by taking the arithmetic inverse of each coefficient (dashed line). The log axis for the magnitudes correctly protrays the ratios.



The filter coefficients for the FIR implementation H(z) can be found by using the inverse FFT:

$LPCINV1 := ifft(cinv)$



6.2 Principal components analysis with Mathcad

Example of Principal Components Analysis with data from Table 1.4.1 of Mardia, Kent and Bibby, 1979.

Input data set X:

$$D1 := \begin{bmatrix} 72 & 66 \\ 60 & 53 \\ 56 & 57 \\ 41 & 29 \\ 32 & 32 \\ 30 & 35 \\ 39 & 39 \\ 42 & 43 \\ 37 & 40 \\ 33 & 29 \\ 32 & 30 \\ 63 & 45 \\ 54 & 46 \\ 47 & 51 \\ 91 & 79 \\ 56 & 68 \\ 79 & 65 \\ 81 & 80 \\ 78 & 55 \\ 46 & 38 \\ 39 & 35 \\ 32 & 30 \\ 60 & 50 \\ 35 & 37 \\ 39 & 36 \\ 50 & 34 \\ 43 & 37 \\ 48 & 54 \end{bmatrix} \quad D2 := \begin{bmatrix} 76 & 77 \\ 66 & 63 \\ 64 & 58 \\ 36 & 38 \\ 35 & 36 \\ 34 & 26 \\ 31 & 27 \\ 31 & 25 \\ 31 & 25 \\ 27 & 36 \\ 34 & 28 \\ 74 & 63 \\ 60 & 52 \\ 52 & 43 \\ 100 & 75 \\ 47 & 50 \\ 70 & 61 \\ 68 & 58 \\ 67 & 60 \\ 37 & 38 \\ 34 & 37 \\ 30 & 32 \\ 67 & 54 \\ 48 & 39 \\ 39 & 31 \\ 37 & 40 \\ 39 & 50 \\ 57 & 43 \end{bmatrix}$$

$$X := \text{augment}(D1, D2) \qquad One := \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Calculate vector of column means:

$$u := \frac{1}{28} \cdot X^T \cdot One \qquad u^T = (\ 50.536 \quad 46.179 \quad 49.679 \quad 45.179\ )$$

Calculate covariance matrix:

$$C := \frac{1}{28} \cdot X^T \cdot X - u \cdot u^T \qquad C = \begin{bmatrix} 280.034 & 215.761 & 278.136 & 218.19 \\ 215.761 & 212.075 & 220.879 & 165.254 \\ 278.136 & 220.879 & 337.504 & 250.272 \\ 218.19 & 165.254 & 250.272 & 217.932 \end{bmatrix}$$

Calculate eigenvalue and assemble matrix (ordered so that eigenvalues are in descending order!):

$$E := \text{eigenvals}(C) \qquad V := \begin{bmatrix} E_3 & 0 & 0 & 0 \\ 0 & E_2 & 0 & 0 \\ 0 & 0 & E_0 & 0 \\ 0 & 0 & 0 & E_1 \end{bmatrix} \qquad V = \begin{bmatrix} 949.283 & 0 & 0 & 0 \\ 0 & 57.651 & 0 & 0 \\ 0 & 0 & 23.058 & 0 \\ 0 & 0 & 0 & 17.553 \end{bmatrix}$$

Calculate eigenvectors and assemble matrix:

$$V0 := \text{eigenvec}\left(C, E_3\right) \qquad V1 := \text{eigenvec}\left(C, E_2\right) \qquad V2 := \text{eigenvec}\left(C, E_0\right) \qquad V3 := \text{eigenvec}\left(C, E_1\right)$$

$$P := \text{augment}(V0, V1) \qquad P := \text{augment}(P, V2) \qquad P := \text{augment}(P, V3)$$

$$P = \begin{bmatrix} 0.526 & -0.225 & -0.678 & -0.462 \\ 0.429 & -0.752 & 0.32 & 0.385 \\ 0.579 & 0.379 & 0.599 & -0.403 \\ 0.452 & 0.49 & -0.282 & 0.69 \end{bmatrix}$$

Verify that Jordans spectral theorem is valid for this example:

$$P2 := P^T \cdot C \cdot P \qquad P2 = \begin{bmatrix} 949.283 & 0 & 0 & 0 \\ 0 & 57.651 & 0 & 0 \\ 0 & 0 & 23.058 & 0 \\ 0 & 0 & 0 & 17.553 \end{bmatrix}$$

Calculate the PCA scores (Y):

$$U := \left(u \cdot (\text{One})^T\right)^T \qquad Y := (X - U) \cdot P$$

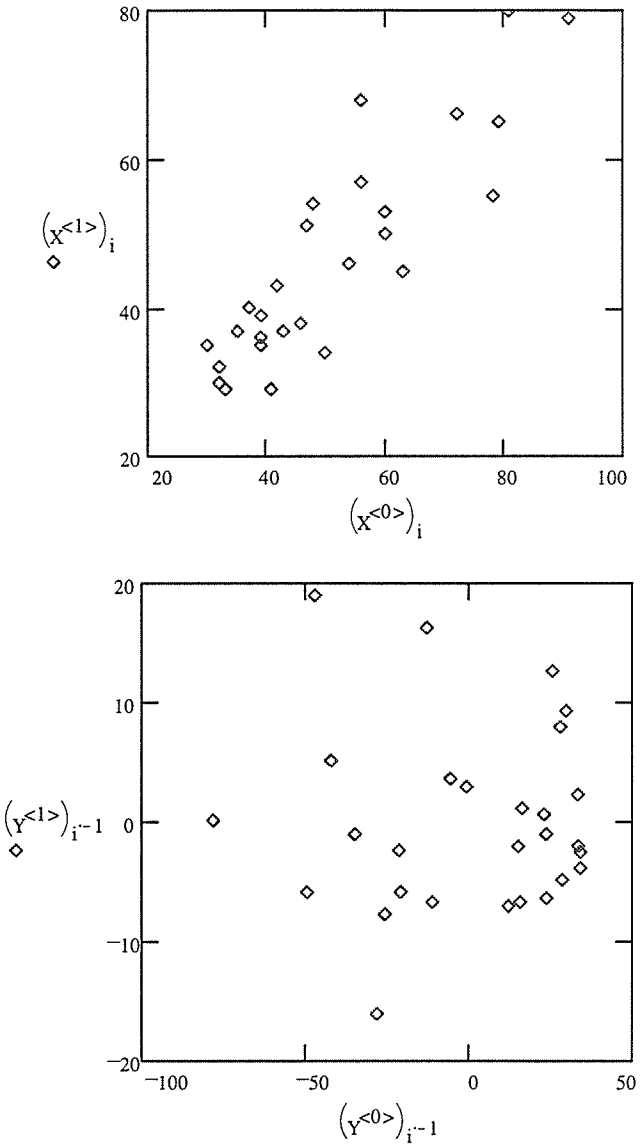
Calculate proportion of variance accounted for by each PCA axis:

$$\text{Var_prop} := \frac{1}{\text{tr}(V)} \cdot \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}^T \cdot V \qquad \text{Var_prop}^T = \begin{bmatrix} 0.906 \\ 0.055 \\ 0.022 \\ 0.017 \end{bmatrix}$$

$$Y = \begin{bmatrix} 49.41 & 5.832 & -1.409 & 9.038 \\ 25.407 & 7.658 & 0.52 & 3.96 \\ 21.6 & 2.34 & 4.723 & 4.705 \\ -23.545 & 6.367 & -5.206 & -1.634 \\ -28.477 & 4.773 & 1.817 & 2.703 \\ -33.341 & -2.315 & 6.355 & -1.715 \\ -28.176 & -7.992 & -0.542 & -2.436 \\ -25.786 & -12.655 & -0.73 & -3.664 \\ -29.703 & -9.275 & 1.697 & -2.507 \\ -33.868 & 3.776 & -4.613 & 4.313 \\ -33.529 & 1.977 & 2.835 & -3.182 \\ 28.186 & 16.031 & 0.719 & -3.729 \\ 10.804 & 6.608 & 1.854 & -1.127 \\ 0.566 & -3.021 & 5.944 & 1.049 \\ 77.97 & -0.109 & 4.819 & -5.803 \\ 12.859 & -16.294 & 0.317 & 10.272 \\ 41.96 & -5.103 & -5.552 & -3.199 \\ 46.93 & -19.064 & -2.458 & 0.381 \\ 34.958 & 1.018 & -9.591 & -6.062 \\ -16.477 & -1.148 & -5.114 & -0.887 \\ -23.634 & 1.055 & -2.847 & 1.714 \\ -34.036 & 2.424 & -0.689 & 1.191 \\ 20.632 & 5.882 & 2.698 & -3.805 \\ -15.873 & 6.731 & 8.326 & 0.065 \\ -23.023 & -0.745 & 2.162 & -4.057 \\ -15.183 & 1.942 & -9.669 & -2.895 \\ -11.903 & 6.917 & -5.589 & 7.586 \\ 5.273 & -3.61 & 9.222 & -0.276 \end{bmatrix}$$

Scattergrams for original axes 0,1 and PCA axes 0,1:

$i := 0..27$



6.3 Tukey Box Plots of feature variables

The eight plots that follow are box plots for the eight variables used in Ch. 3, as they are output by the SAS (1985a) procedure UNIVARIATE. These plots follow the format outlined by Tukey (1977). The following characteristics of the data are displayed in these plots:

a) 25th percentile (bottom of the box),

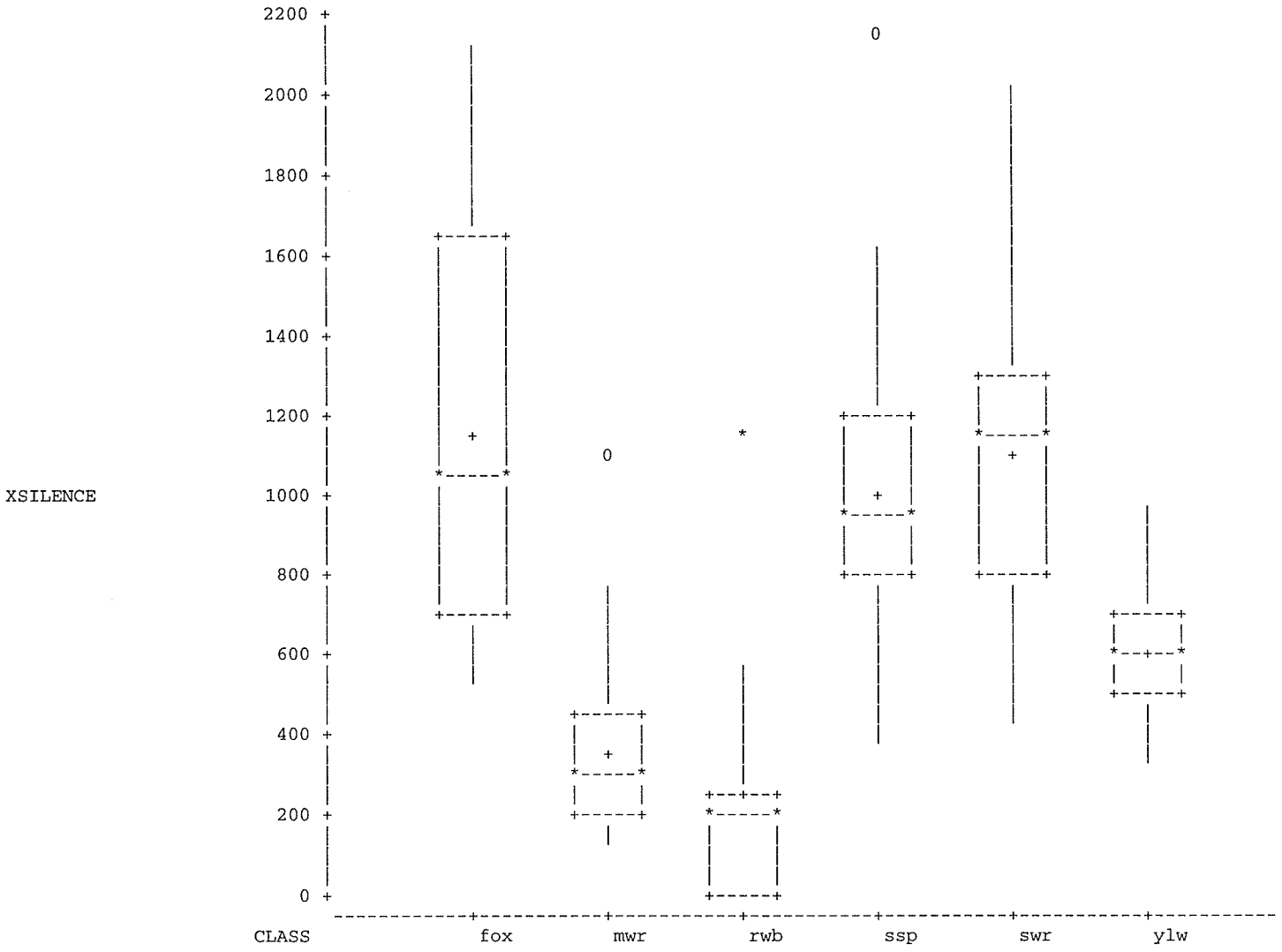
b) 75th percentile (top of the box),

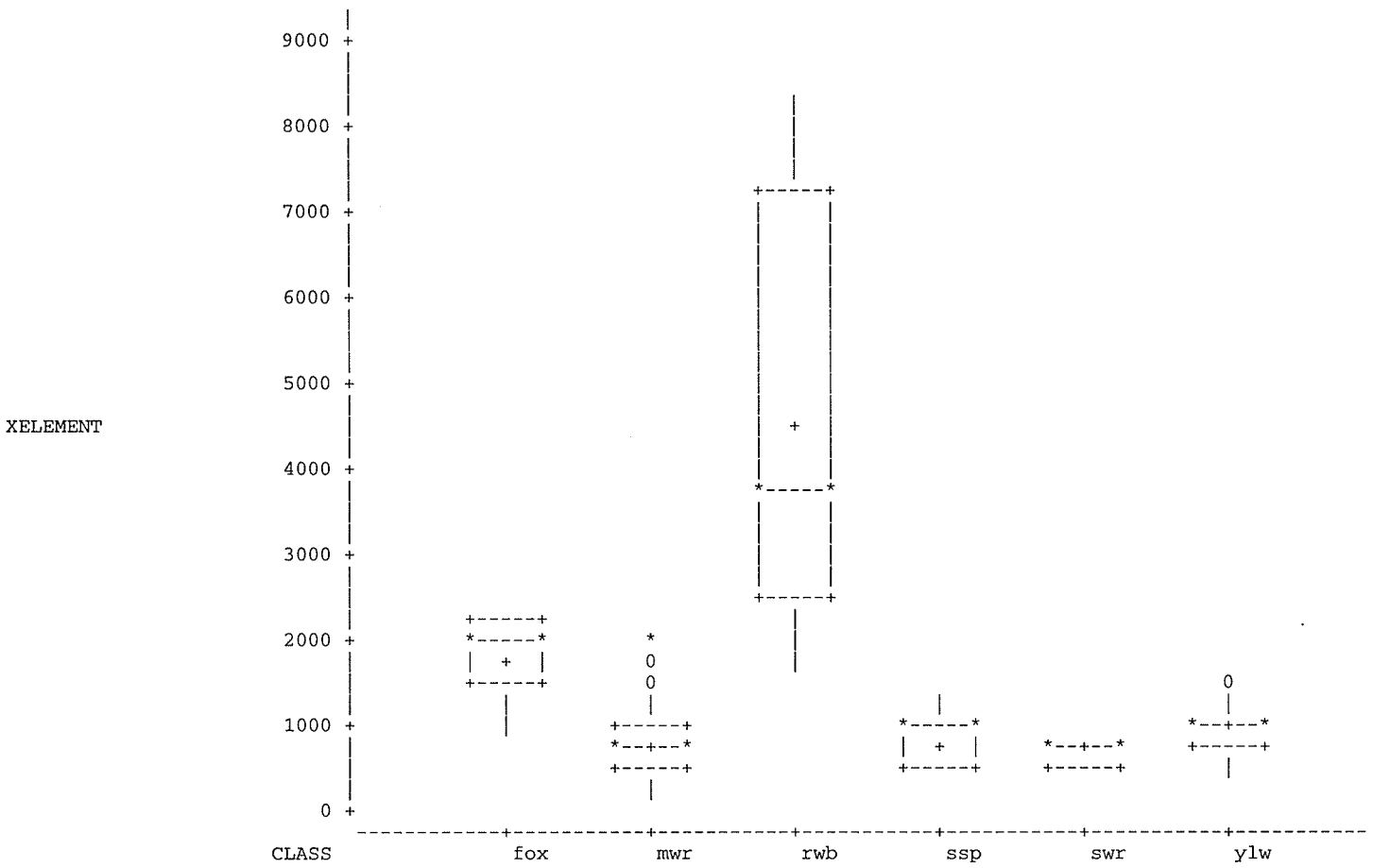
c) median (horizontal line within the box),

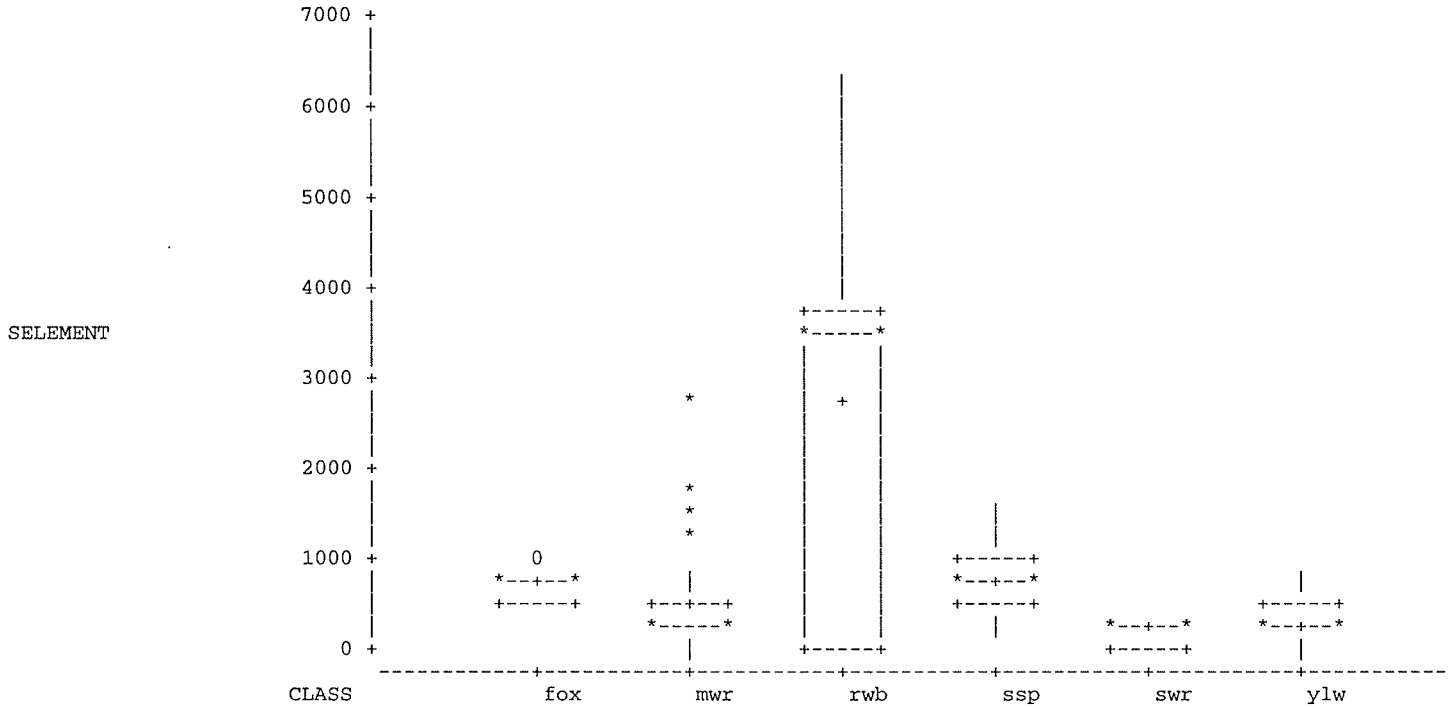
d) mean (central '+' sign),

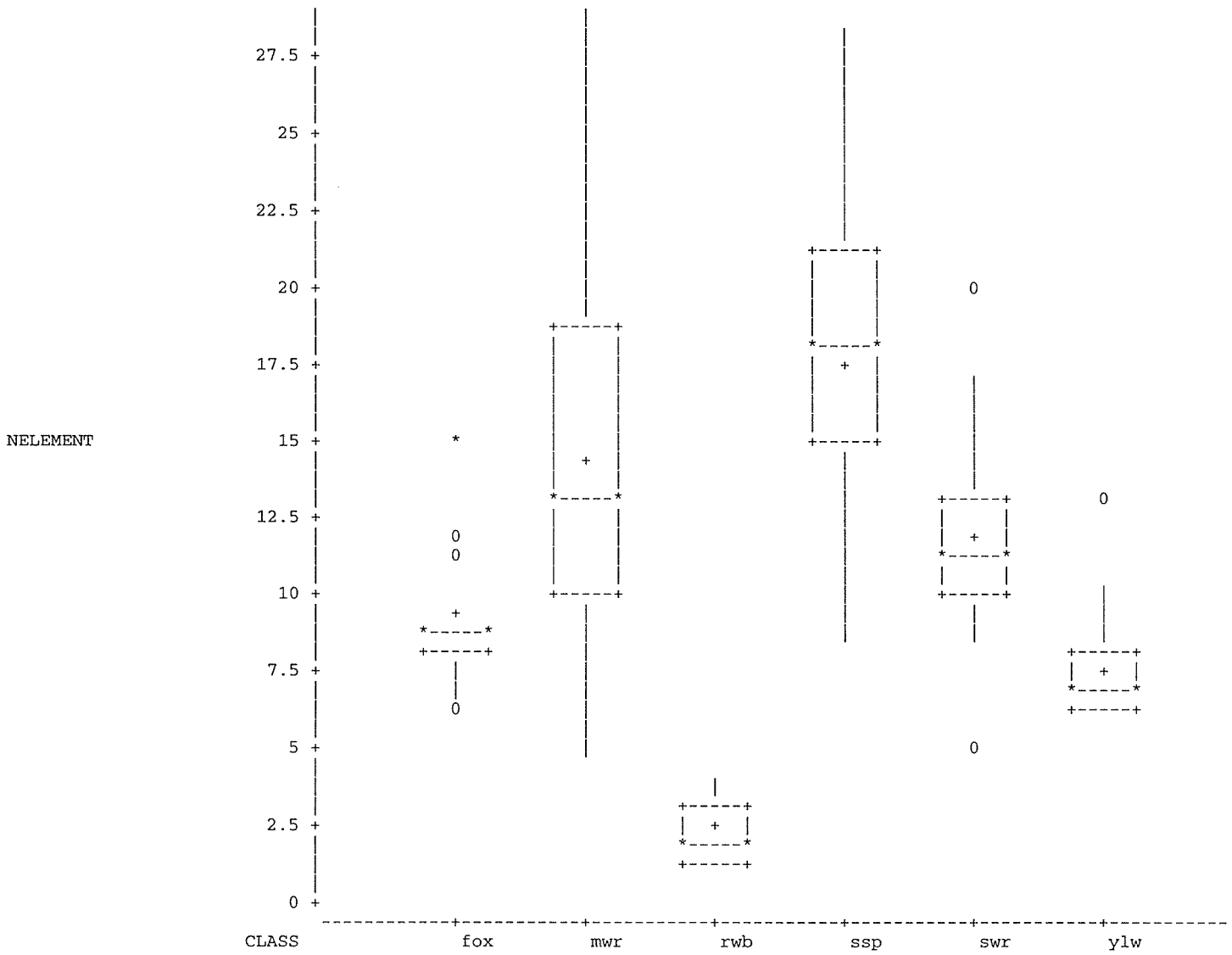
e) the vertical 'whiskers' extend from the box to 1.5 time the distance between the 25th and 75th percentile (interquartile range) if there is data in this range,

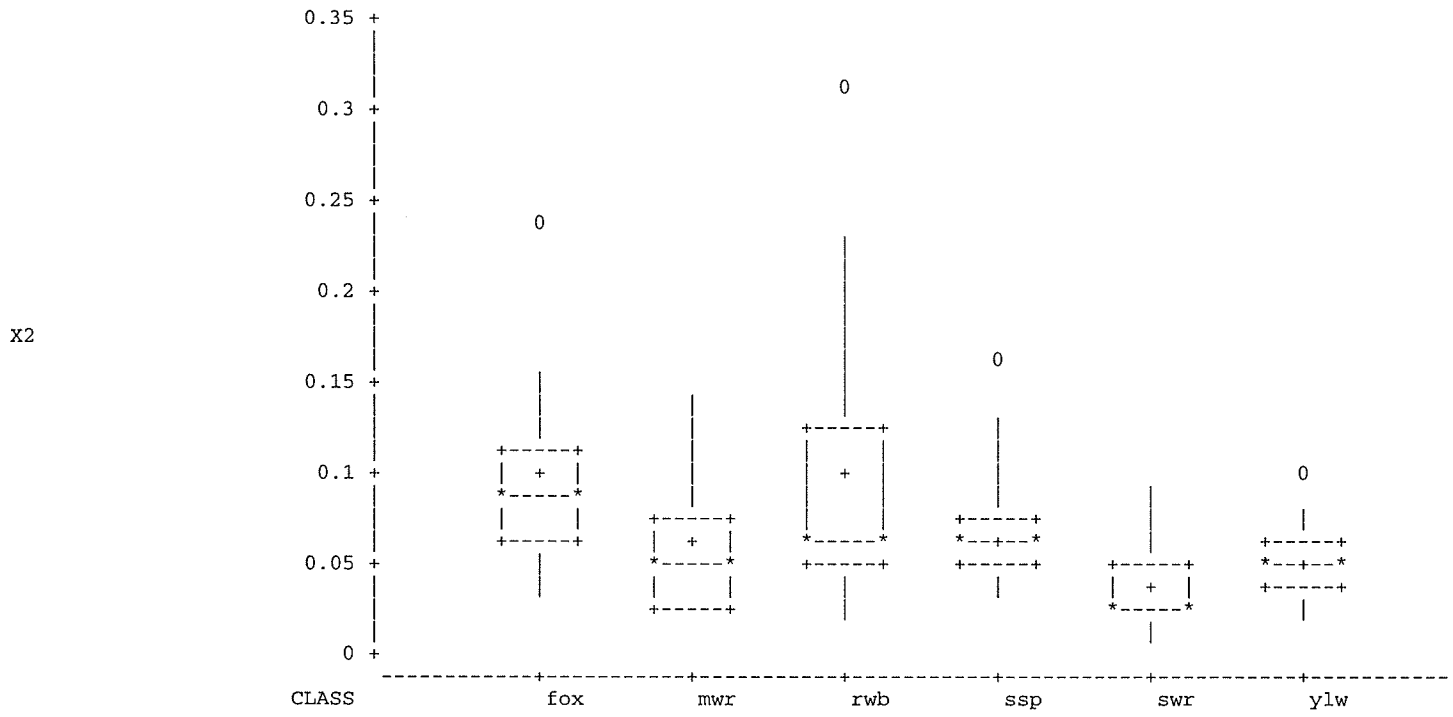
f) more extreme values than those indicated by whiskers are shown as zeros if they are within 3 times the interquartile range or as asterisks if they are even more extreme than this.

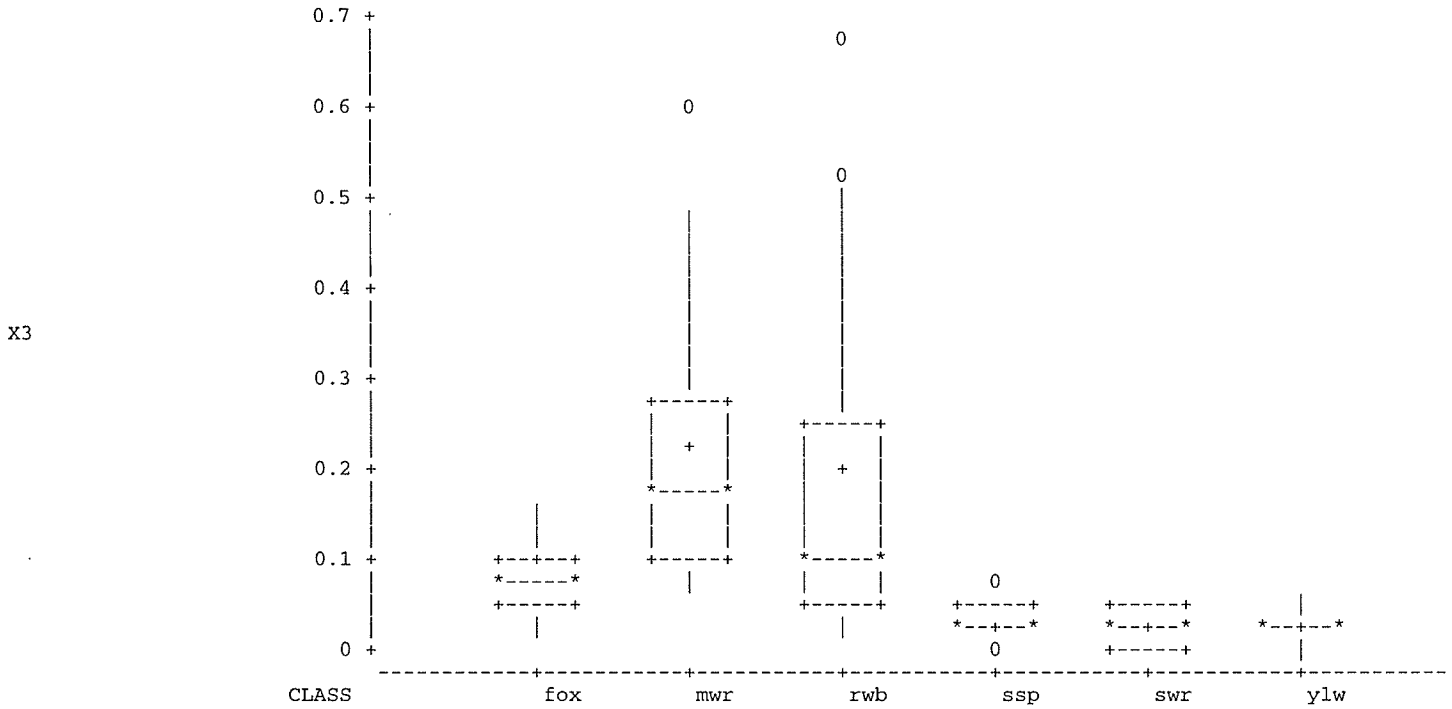

XSILENCE
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
CLASS
fox
mwr
rwb
ssp
swr
ylw

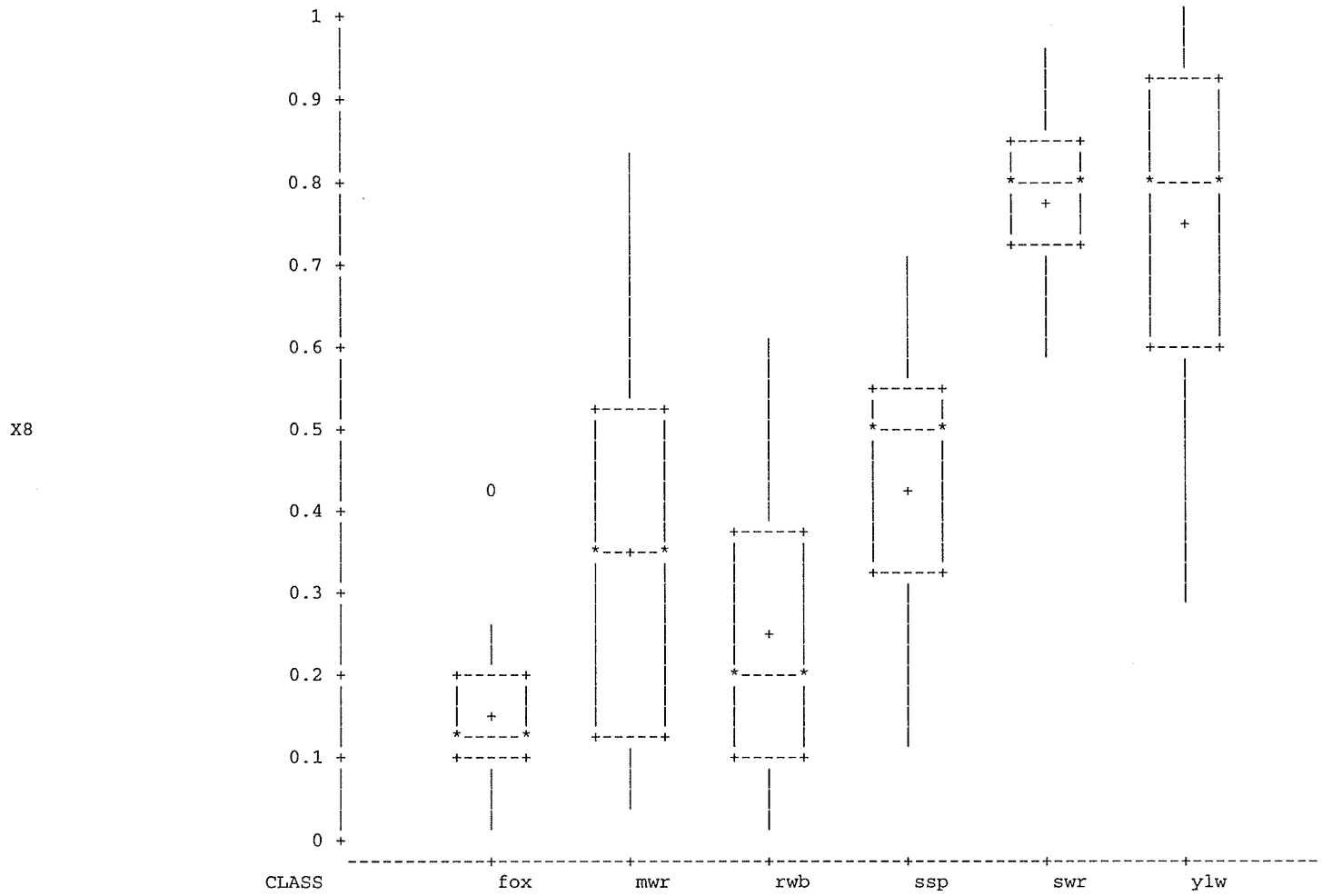

XELEMENT
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
CLASS
fox
mwr
rwb
ssp
swr
ylw

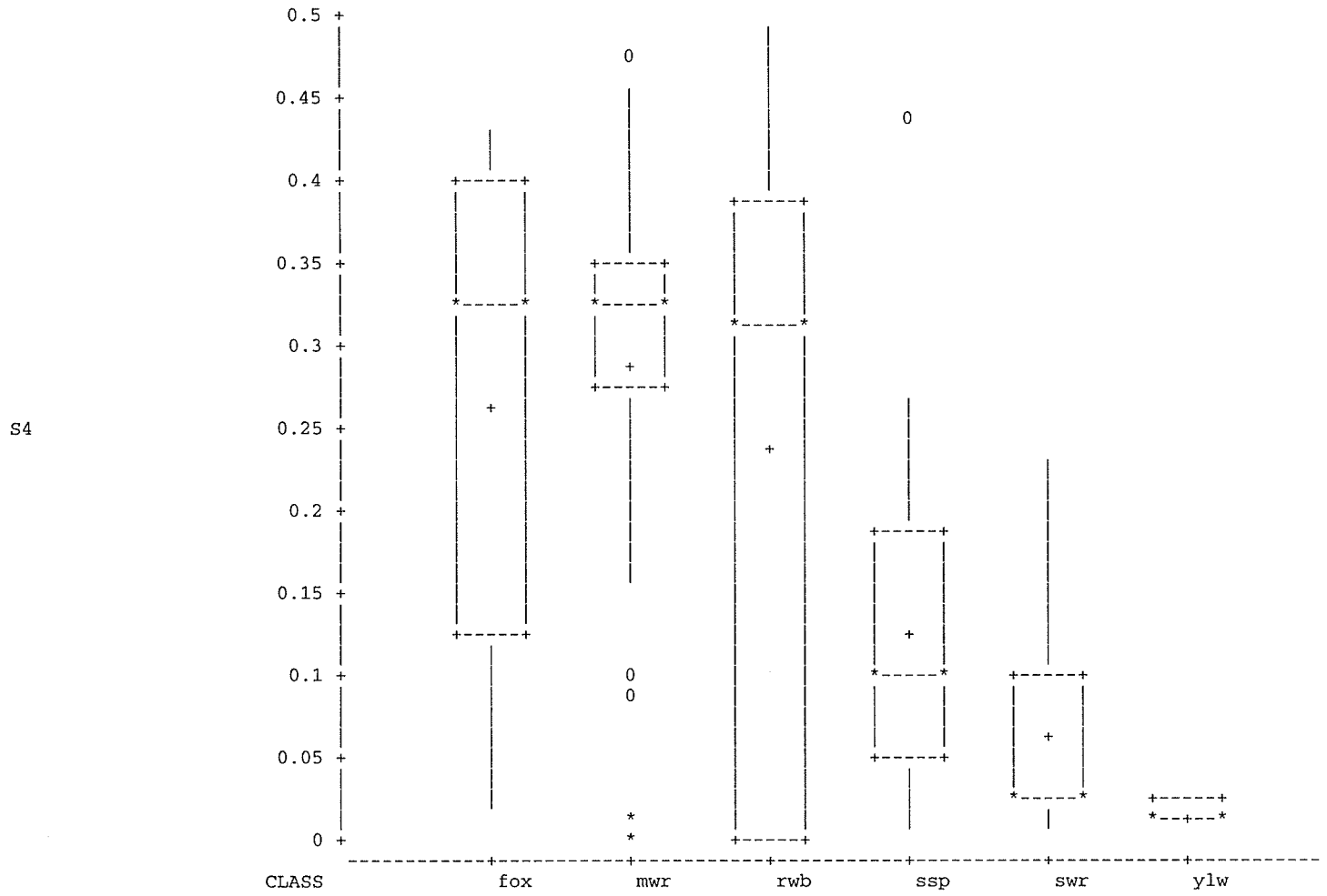

SELEMENT
7000
6000
5000
4000
3000
2000
1000
0
CLASS
fox
mwr
rwb
ssp
swr
ylw


NELEMENT
27.5
25
22.5
20
17.5
15
12.5
10
7.5
5
2.5
0
CLASS
fox
mwr
rwb
ssp
swr
ylw


X2
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
CLASS
fox
mwr
rwb
ssp
swr
ylw


X3
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
CLASS
fox
mwr
rwb
ssp
swr
ylw


X8
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
CLASS
fox
mwr
rwb
ssp
swr
ylw


S4
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
CLASS
fox
mwr
rwb
ssp
swr
ylw