

License granted by Englehart, Kevin on 2006-09-26T19:06:08Z (GMT) :

===== | Use Licence | =====

Attribution-NonCommercial-ShareAlike 1.0

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works

Under the following conditions:

- Attribution.

You must give the original author credit.

- Non-Commercial.

You may not use this work for commercial purposes.

- Share Alike.

If you alter, transform, or build upon this work, you may distribute the resulting work only under a license identical to this one.

For any reuse or distribution, you must make clear to others the license terms of this work. Any of these conditions can be waived if you get permission from the author.

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

This work is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit:

URL (human-readable summary): <http://creativecommons.org/licenses/by-nc-sa/1.0/>
URL (legal code): <http://creativecommons.org/worldwide/uk/translated-license>

=====

===== | Site Licence | =====

This is the default license that MIT requires all submitters to grant.
It is provided for informational purposes only.

NON-EXCLUSIVE DISTRIBUTION LICENSE

By signing and submitting this license, you (the author(s) or copyright owner) grants to Massachusetts Institute of Technology (MIT) the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video.

You agree that MIT may, without changing the content, translate the submission to any medium or format for the purpose of preservation.

You also agree that MIT may keep more than one copy of this submission for purposes of security, back-up and preservation.

You represent that the submission is your original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright.

If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant MIT the rights required by this license, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission.

IF THE SUBMISSION IS BASED UPON WORK THAT HAS BEEN SPONSORED OR SUPPORTED BY AN AGENCY OR ORGANIZATION OTHER THAN MIT, YOU REPRESENT THAT YOU HAVE FULFILLED ANY RIGHT OF REVIEW OR OTHER OBLIGATIONS REQUIRED BY SUCH CONTRACT OR AGREEMENT.

MIT will clearly identify your name(s) as the author(s) or owner(s) of the submission, and will not make any alteration, other than as allowed by this license, to your submission.

=====

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI[®]

NOTE TO USERS

This reproduction is the best copy available

UMI

**SIGNAL REPRESENTATION FOR CLASSIFICATION
OF THE TRANSIENT MYOELECTRIC SIGNAL**

by

Kevin Brian Englehart

B.Sc. (E.E.), University of New Brunswick, 1989
M.Sc. (E.E.), University of New Brunswick, 1991

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of

Doctor of Philosophy

in the Department of Electrical and Computer Engineering

Supervisors: Philip Parker, Electrical and Computer Engineering
Maryhelen Stevenson, Electrical and Computer Engineering

Examining Board: Bernard Hudgins, Institute of Biomedical Engineering
Dana Wasson, Computer Science
Huw Davies, Mechanical Engineering.

External Examiner: Roberto Merletti, Politecnico di Torino, Italy

THE UNIVERSITY OF NEW BRUNSWICK

October, 1998

© Kevin Englehart, 1998



National Library
of Canada

Acquisitions and
Bibliographic Services

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale
du Canada

Acquisitions et
services bibliographiques

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

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

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

0-612-46463-6

Canada

Abstract

The myoelectric signal (MES) is the electrical manifestation of muscular contraction. The MES, recorded at the surface of the skin, has been exploited as a control source for powered upper extremity prosthetics. The primary objective of this work is to improve the accuracy with which transient patterns of MES activity may be classified, and to do so in a computationally efficient manner. The direct motivation of this work is to enhance a myoelectric control scheme based upon classification of these patterns.

The classification problem may be divided into the stages of feature extraction, dimensionality reduction, and pattern recognition. It is shown in this work that classification performance depends crucially upon the signal representation, which comprises the stages of feature extraction and dimensionality reduction. It is shown that linear time-frequency representations (the short-time Fourier transform, the wavelet transform, and the wavelet packet transform) provide a powerful framework for feature extraction, localizing the discriminant information of the transient MES in time and in frequency. Due to the high dimension of linear time-frequency representations, their success relies upon an appropriate form of dimensionality reduction. It is shown that principal components analysis (PCA) provides an effective means of concentrating that information which is important, and discarding that which is irrelevant. A wavelet packet based feature set, subject to PCA, is shown to outperform all other forms of signal representation.

An assessment of the classification performance of the wavelet packet / PCA signal representation suggests that it may be yielding near-optimal results (the Bayes bound). The improved accuracy provided by these methods will enhance

the functionality of a pattern recognition based myoelectric control system. The complexity of this signal representation is on the order of $N \log N$, easily lending itself to real-time implementation on a DSP microprocessor of modest capabilities. Moreover, the task of class separation has been accommodated by the signal representation, allowing a simple classifier (a linear discriminant analysis) to be used in lieu of a neural network. This obviates the need for fine tuning the architecture and the training algorithm of the classifier, which is beneficial in clinical applications.

Table of Contents

| | |
|---|-------------|
| Abstract | ii |
| List of Figures | ix |
| List of Tables..... | xii |
| List of Symbols and Abbreviations..... | xiii |
| Acknowledgements | xv |
| | |
| Chapter 1 – Introduction..... | 1 |
| 1.1 Objectives | 1 |
| 1.2 Myoelectric Control | 3 |
| 1.3 The Transient Myoelectric Signal..... | 7 |
| 1.3.1 The Etiology of the MES | |
| 1.3.2 The MES During Sustained Contractions | |
| 1.3.3 Structure in the Transient MES | |
| 1.3.4 Two Channel Transient MES Patterns | |
| 1.4 Thesis Outline..... | 14 |
| | |
| Chapter 2 – Signal Representation for Classification | 16 |
| 2.1 Introduction | 16 |
| 2.2 Problem Formulation..... | 17 |
| 2.3 Pattern Classification..... | 20 |
| 2.3.1 Bayesian Pattern Classification | |
| 2.3.1.1 The Gaussian Bayes Classifier | |
| 2.3.2 Artificial Neural Networks | |
| 2.3.2.1 The Perceptron | |
| 2.3.2.2 The Multilayer Perceptron | |
| 2.3.2.3 Issues in MLP ANN Training | |
| 2.3.2.4 The Capabilities of ANNs as Classifiers | |
| 2.3.3 Other Classifiers | |
| 2.3.3.1 Distance Classifiers | |
| 2.3.3.2 Classification and Regression Trees | |
| 2.3.4 Summary: Pattern Classification | |
| 2.4 Feature Extraction..... | 44 |
| 2.4.1 The Importance of Feature Extraction | |
| 2.4.2 Feature Extraction for Classification | |

| | |
|---|------------|
| 2.5 Time-Frequency Representations | 53 |
| 2.5.1 The Short-Time Fourier Transform | |
| 2.5.2 The Wavelet Transform | |
| 2.5.3 Quadratic Time-Frequency Representations | |
| 2.5.4 Time-Frequency Representations for Classification | |
| 2.6 Dimensionality Reduction..... | 74 |
| 2.6.1 Feature Selection | |
| 2.6.2 Feature Projection | |
| 2.6.2.1 Principal Components Analysis | |
| 2.6.2.2 Projection Pursuit | |
| 2.6.2.3 Neural Networks | |
| 2.6.3 Summary: Dimensionality Reduction | |
| 2.7 Summary..... | 93 |
| | |
| Chapter 3 – Wavelet-Based Feature Extraction | 95 |
| 3.1 Introduction..... | 95 |
| 3.2 Wavelet Bases..... | 99 |
| 3.2.1 The Continuous Wavelet Transform | |
| 3.2.2 The Discrete Wavelet Transform | |
| 3.2.3 Multiscale Filter Banks | |
| 3.2.4 A Simple Example | |
| 3.2.5 Time-Frequency Plane Tiling | |
| 3.2.6 Wavelet Selection | |
| 3.3 Wavelet Packet and Cosine Packet Bases | 115 |
| 3.3.1 The Wavelet Packet Transform | |
| 3.3.2 The Cosine Packet Transform | |
| 3.4 Best Basis Selection | 126 |
| 3.4.1 Best Bases for Signal Compression | |
| 3.4.2 Best Bases for Classification | |
| 3.4.2.1 Discriminant Measures | |
| 3.4.2.2 The Local Discriminant Basis Algorithm | |
| 3.4.2.3 Best Bases for Classification: Some Examples | |
| 3.5 Summary..... | 145 |

| | |
|--|------------|
| Chapter 4 – Time-Frequency Methods for MES Classification | 146 |
| 4.1 Preliminary Issues..... | 147 |
| 4.1.1 The Transient MES Data | |
| 4.1.2 Real-Time Considerations | |
| 4.1.3 Methodology | |
| 4.1.3.1 Datasets | |
| 4.1.3.2 Feature Sets | |
| 4.1.3.3 Dimensionality Reduction | |
| 4.1.3.4 Classifiers | |
| 4.1.4 Dimensionality Reduction via Validation | |
| 4.1.5 Summary | |
| 4.2 Time Domain Features | 166 |
| 4.2.1 Derivation of the Time Domain Set | |
| 4.2.2 Feature Set Parameters | |
| 4.2.2.1 Record Length | |
| 4.2.2.2 Waveform Segmentation | |
| 4.2.3 Dimensionality Reduction | |
| 4.2.3.1 Feature Selection | |
| 4.2.3.2 Optimal Segmentation with Feature Selection | |
| 4.2.3.3 Optimal Segmentation with Feature Projection | |
| 4.2.4 Summary | |
| 4.3 STFT Based Features | 191 |
| 4.3.1 Related Work | |
| 4.3.2 STFT Feature Extraction Parameters | |
| 4.3.2.1 STFT Window Size | |
| 4.3.2.2 STFT Window Overlap | |
| 4.3.2.3 STFT Window Type | |
| 4.3.3 Dimensionality Reduction Methods | |
| 4.3.3.1 Feature Selection | |
| 4.3.3.2 Feature Projection | |
| 4.3.3.3 Bin Averaging | |
| 4.3.3.4 Relative Performance of Dimensionality Reduction Strategies | |
| 4.3.4 Summary | |
| 4.4 Wavelet Transform Based Features..... | 217 |
| 4.4.1 Background | |
| 4.4.2 Related Work | |

| | | |
|---|--|------------|
| 4.4.3 | Wavelet Transform Parameter Selection | |
| 4.4.3.1 | The Mother Wavelet | |
| 4.4.3.2 | The Depth of Decomposition | |
| 4.4.4 | Dimensionality Reduction | |
| 4.4.4.1 | Feature Selection | |
| 4.4.4.2 | Feature Projection | |
| 4.4.4.3 | Wavelet Transform Local Extrema | |
| 4.4.4.4 | Wavelet Transform Subband Energy | |
| 4.4.4.5 | The Relative Performance of Dimensionality Reduction Methods | |
| 4.4.5 | Summary | |
| 4.5 | Wavelet Packet Based Features | 238 |
| 4.5.1 | Related Work | |
| 4.5.2 | Wavelet Transform Parameter Selection | |
| 4.5.2.1 | Selection of the Mother Wavelet and LDB Cost Function | |
| 4.5.2.2 | A Temporally Segmented LDB | |
| 4.5.3 | Dimensionality Reduction | |
| 4.6 | Performance Summary | 252 |
| 4.6.1 | The Relative Performance Amongst Feature Sets | |
| 4.6.2 | Summary | |
| Chapter 5 – Assessment of Classification Performance | | 263 |
| 5.1 | Generalized Feature Dimension | 264 |
| 5.2 | Performance Bounds | 272 |
| 5.2.1 | Hybrid Feature Sets | |
| 5.2.2 | Two-Class LDB Formulation | |
| 5.2.3 | Bootstrapping with Noise | |
| 5.2.4 | Spin Cycle Training | |
| 5.2.5 | Summary | |
| 5.3 | Modeling the Transient MES Problem | 288 |
| 5.3.1 | Additive Noise | |
| 5.3.2 | Temporal Translation | |
| 5.3.3 | Superposition of Motor Groups | |
| 5.3.4 | Wavelet Transform Decomposition Model | |
| 5.3.4.1 | Modeling the Intra-Class Variance | |
| 5.3.4.2 | Bootstrapping using Simulated MES Data | |

| | |
|---|------------|
| 5.4 Summary..... | 308 |
| | |
| Chapter 6 – Conclusions | 310 |
| 6.1 Summary..... | 310 |
| 6.2 Original Contributions | 315 |
| 6.3 Future Work..... | 317 |
| | |
| References | |
| | |
| Appendix A – The Backpropagation Algorithm | |
| | |
| Appendix B – Quadratic Time-Frequency Representations | |
| | |
| Appendix C – Time-Frequency Plane Tiling of Wavelet and Cosine Packet Transforms | |
| | |
| Appendix D – The Lack of Shift Invariance of the Wavelet Transform | |

List of Figures

| | |
|---|-----|
| Figure 1.1 – Hudgins' multifunction control scheme. | 5 |
| Figure 1.2 – A single motor unit. | 7 |
| Figure 1.3 – Single Channel MES activity recorded from the biceps and triceps. | 10 |
| Figure 1.4 – Two Channel MES activity recorded from the biceps and triceps. | 13 |
| Figure 2.1 – The Classification Problem. | 20 |
| Figure 2.2 – A Bayesian Classifier | 24 |
| Figure 2.3 – Artificial neural networks that have been used as pattern classifiers. | 28 |
| Figure 2.4 – The Perceptron. | 31 |
| Figure 2.5 – Activation functions in an artificial neuron. | 32 |
| Figure 2.6 – The architecture of a typical MLP network. | 33 |
| Figure 2.7 – Distance classifiers. | 39 |
| Figure 2.8 – A refined model of the classification problem | 51 |
| Figure 2.9 – Dirac and Fourier bases | 57 |
| Figure 2.10 – The spectrogram of two sinusoids and two delta functions. | 57 |
| Figure 2.11 – The tiling in the time-frequency plane of the STFT. | 59 |
| Figure 2.12 – A 256-point linear chirp signal. | 60 |
| Figure 2.13 – Three discrete STFT representations of a linear chirp signal. | 60 |
| Figure 2.14 – Division of the frequency domain for the STFT and the WT. | 63 |
| Figure 2.15 – The time-frequency plane tiling of the WT and WPT | 64 |
| Figure 2.16 – The TFRs of wavelet-based representations. | 65 |
| Figure 2.17 – The TFRs of three localized sinusoids. | 69 |
| Figure 2.18 – The TFRs of MES activity. | 72 |
| Figure 2.19 – The processing stages of classification: notation. | 75 |
| Figure 2.20 – Feature extraction for signal compression vs. signal classification. | 87 |
| Figure 2.21 – An auto-associative MLP having two layers of weights. | 90 |
| Figure 3.1 – Some symmetric wavelets at various scales and locations. | 100 |
| Figure 3.2 – Subband decomposition analogy of the wavelet transform | 103 |
| Figure 3.3 – Projection operators H and G . | 104 |
| Figure 3.4 – The subband coding analogy of the DWT. | 105 |
| Figure 3.5 – The Daubechies-4 wavelet and scaling functions | 107 |
| Figure 3.6 – The 3-scale decomposition of a chirp signal. | 108 |
| Figure 3.7 – The tiling of the time-scale domain (a) | 110 |
| Figure 3.8 – The 3-scale DWT time-frequency response of a chirp signal. | 111 |
| Figure 3.9 – The time-frequency plane tiling of the WT and WPT | 116 |
| Figure 3.10 – The first 2 levels of decomposition in a wavelet packet transform. | 116 |
| Figure 3.11 – The binary wavelet packet tree for a Kronecker delta function | 117 |
| Figure 3.12 – The binary wavelet packet tree for a sinusoid of $N=32$ samples | 118 |
| Figure 3.13 – The binary wavelet packet tree for a 256-sample chirp signal. | 118 |
| Figure 3.14 – A decomposition into binary tree-structured subspaces | 119 |
| Figure 3.15 – Some wavelet packet basis vectors for scale=5 and scale=6. | 120 |
| Figure 3.16 – The time-frequency plane tiling of wavelet, cosine packet bases | 121 |
| Figure 3.17 – The binary cosine packet tree for a Kronecker delta function. | 122 |
| Figure 3.18 – Some cosine packet basis functions. | 123 |
| Figure 3.19 – A bell function used to achieve temporal localization. | 123 |

| | |
|--|-----|
| Figure 3.20 – The characteristics of the wavelet packet best basis. | 130 |
| Figure 3.21 – The characteristics of the cosine packet best basis. | 131 |
| Figure 3.22 – Reconstruction of a chirp signal using the WT, WPT, and CPT. | 132 |
| Figure 3.23 – The compression of a transient MES pattern | 132 |
| Figure 3.24 – An elbow flexion pattern subject to WPT compression. | 133 |
| Figure 3.25 – Packet decomposition of a training set into energy maps, | 133 |
| Figure 3.26 – Noisy chirp signals | 141 |
| Figure 3.27 – The LDB's discriminant measures in time-frequency plane. | 142 |
| Figure 3.28 – A scatterplot of the coefficients of 24 LDB bases. | 143 |
| Figure 3.29 – Classification of noisy chirp signals (SNR=0.5) | 144 |
| Figure 4.1 – Datasets, feature extraction, dimensionality reduction, and classifiers. | 151 |
| Figure 4.2 – Scatterplots of the CS-ranked STFT coefficients | 156 |
| Figure 4.3 – Scatterplots of the PCA-ranked STFT coefficients | 157 |
| Figure 4.4 – Validation based dimensionality specification. | 164 |
| Figure 4.5 – Waveform segmentation for time domain feature extraction. | 167 |
| Figure 4.6 – One channel data: the effect of record length | 172 |
| Figure 4.7 – Two channel data: the effect of record length. | 173 |
| Figure 4.8 – One channel data: all segmentation schemes | 175 |
| Figure 4.9 – The effect of the number of segments in a $N=240$ record | 177 |
| Figure 4.10 – The effect of omitting MAVS and SSC. | 179 |
| Figure 4.11 – Time domain feature ranking for Subject 5 | 181 |
| Figure 4.12 – Feature ranking (all subjects) using class separability | 182 |
| Figure 4.13 – Subsets of time domain features: CS, AO and KO methods. | 183 |
| Figure 4.14 – The effect of record length with feature selection | 184 |
| Figure 4.15 – The effect of segmentation with feature selection | 185 |
| Figure 4.16 – The effect of record length with PCA feature projection | 186 |
| Figure 4.17 – The effect of segmentation with PCA feature projection. | 187 |
| Figure 4.18 – Dimensionality reduction vs. waveform segmentation. | 189 |
| Figure 4.19 – Optimal performance of time domain features | 190 |
| Figure 4.20 – The effect of STFT window size upon classification error. | 197 |
| Figure 4.21 – The effect of STFT window overlap upon classification error | 199 |
| Figure 4.22 – The envelope and frequency response of STFT windows. | 201 |
| Figure 4.23 – The effect of STFT window type upon classification error | 203 |
| Figure 4.24 – The performance of CS, AO, KO vs STFT feature dimension | 206 |
| Figure 4.25 – The performance of CS, AO, KO at optimal dimension..... | 207 |
| Figure 4.26 – The performance of PCA vs STFT feature dimension. | 208 |
| Figure 4.27 – Performance of CS, PCA, and bin averaging with STFT vs. dimension.... | 213 |
| Figure 4.28 – Performance of CS, PCA, and bin averaging with STFT at optimal. dimension.. | 215 |
| Figure 4.29 – The classification error when using various types of mother wavelet | 227 |
| Figure 4.30 – The time-frequency tiling of a full and partial WT decomposition. | 229 |
| Figure 4.31 – The effect of the level of WT decomposition | 229 |
| Figure 4.32 – CS-reduced TD, STFT and WT feature sets vs. dimension | 231 |
| Figure 4.33 –PCA-reduced TD, STFT and WT feature sets vs. dimension. | 232 |
| Figure 4.34 – Performance of the WT using CS, PCA, LE and SE dimensionality reduction.... | 236 |
| Figure 4.35 – Wavelet families, when using an I-divergence LDB cost function. | 242 |
| Figure 4.36 – Wavelet families, when using a J-divergence LDB cost function. | 243 |
| Figure 4.37 – Wavelet families, when using an Euclidean distance LDB cost function. .. | 243 |
| Figure 4.38 – Wavelet families, when using an entropy LDB cost function | 244 |
| Figure 4.39 – Relative performance of each LDB cost function. | 245 |

| | |
|---|-----|
| Figure 4.40 – The time-frequency tiling of a temporally segmented LDB | 247 |
| Figure 4.41 – The test set classification error of a temporally segmented LDB | 248 |
| Figure 4.42 – CS-reduced TD, STFT, WT and WPT feature sets vs. dimension. | 250 |
| Figure 4.43 – PCA-reduced TD, STFT, WT and WPT feature sets vs. dimension. | 251 |
| Figure 4.44 – The optimal test set classification error for all feature sets | 253 |
| Figure 4.45 – Scatterplot of the test set classification error, all feature sets | 255 |
| Figure 4.46 – Scatterplot of the normalized test set classification error, all feature sets. | 257 |
| Figure 4.47 – A histogram of the best and worst feature sets. | 260 |
| Figure 5.1 – Validation vs. generalized feature dimension specification | 265 |
| Figure 5.2 – Validation set error vs. feature set dimension. | 266 |
| Figure 5.3 – Validation dimension specification for each feature set | 268 |
| Figure 5.4 – Validation set error vs. feature set dimension averaged across all subjects . | 269 |
| Figure 5.5 – Test set classification error: validation vs. generalized dimension | 270 |
| Figure 5.6 – A hybrid feature set. | 275 |
| Figure 5.7 – The test set classification error when using a hybrid feature set | 275 |
| Figure 5.8 – A classification scheme implementing an ensemble of 2-class LDBs. | 280 |
| Figure 5.9 – The relative performance of 4-class and 2-class LDB schemes | 278 |
| Figure 5.10 – The effect of bootstrapping with noise with CS-reduced features. | 281 |
| Figure 5.11 – The effect of bootstrapping with noise with PCA-reduced features. | 282 |
| Figure 5.12 – The effects of spin cycling upon the test set classification error | 284 |
| Figure 5.13 – The performance of each feature set in additive noise. | 289 |
| Figure 5.14 – The effect of temporal shift upon each feature | 291 |
| Figure 5.15 – Patterns used to represent the activity of individual motor groups | 294 |
| Figure 5.16 – The test set classification error of the superposition-based dataset | 295 |
| Figure 5.17 – Triangular waveforms. | 296 |
| Figure 5.18 – Performance of the triangular waveform vs. feature set dimension. | 297 |
| Figure 5.19 – The reconstruction of the four largest WT coefficients.. | 300 |
| Figure 5.20 – Ensembles of real and simulated transient MES patterns | 302 |
| Figure 5.21 – The effect of the maximum range of shift in the WT model | 303 |
| Figure 5.22 – The effect of the maximum range of amplitude scale in the WT model ... | 303 |
| Figure 5.23 – Performance of the simulated transient MES vs. feature set dimension | 304 |
| Figure 5.24 – The effect of bootstrapping a sparse training set using the WT model | 306 |
| Figure 5.25 – Bootstrapping a sparse training set from a limb-deficient individual.. | 307 |
| Figure B.1 – The TFRs of a time-localized sinusoid | B.2 |
| Figure B.2 – The TFRs of a multicomponent signal | B.4 |
| Figure C.1 – An information cell in the time-frequency plane | C.1 |
| Figure D.1 – The subband coding analogy of the DWT | D.1 |
| Figure D.2 – Output of the WT filter bank, demonstrating aliasing due to decimation .. | D.2 |
| Figure D.3 – The WT coefficients of an unshifted and a shifted MES pattern | D.3 |
| Figure D.4 – The TFR of an unshifted and a shifted MES pattern | D.4 |
| Figure D.5 – The effect of shift-induced dispersion in a CS-reduced WT feature set | D.6 |
| Figure D.6 – The effect of shift-induced dispersion in a PCA-reduced WT feature set ... | D.6 |
| Figure D.7 – Artificial dataset: dispersion of the first three PCA features | D.8 |
| Figure D.8 – Real MES dataset: dispersion of the first ten PCA features | D.9 |
| Figure D.9 – Real MES dataset: effect of shift upon classification performance | D.9 |

List of Tables

| | |
|--|-----|
| Table 4.1 – The combinations of segment length and the number of frames used as possible waveform segmentation schemes for TD analysis | 175 |
| Table 4.2 – The “best” and “worst” features for Subject 5, using all time domain features, evaluated upon six 40 ms segments. | 180 |
| Table 4.3 – Scheffe test results, providing a comparison amongst the performance of all feature sets. | 258 |
| Table 4.4 – The homogeneous subsets of feature sets indicated by the the Scheffe test ($\alpha = 0.05$)..... | 259 |

List of Symbols and Abbreviations

| | |
|---------------------|--|
| \mathcal{X} | The measurement space. This is a time series, such that $\mathbf{x} \in \mathcal{X} \subseteq \Re^N$, where $\mathbf{x} = [x_1, x_2, \dots, x_N]^T$. This is equivalent to the expression $\mathbf{x} = \{x[1], x[2], \dots, x[N]\}$. |
| \mathcal{Y} | The response space. The output may be assigned to one of K classes, $y \in \mathcal{Y} = \{y_1, \dots, y_K\}$. |
| H, G | The lowpass and highpass WT projection operators, including the quadrature mirror filters h and g , and a decimation by two. |
| K | The dimension of the response space (the number of classes). |
| L | The dimension of the reduced feature set. |
| M | The dimension of the original feature set. |
| N | The dimension of the measurement space (the length of the time series in samples). |
| $D(p_i, q_i)$ | A discriminant measure of the i^{th} feature between classes p and q . |
| $O(N)$ | Computational complexity on the order of N . |
| $\Psi_{j,n}$ | The mother wavelet at scale j and location n . |
| $\phi_{j,n}$ | The scaling function at scale j and location n . |
| $D_j[n]$ | The WT detail coefficient at scale j and location n . |
| $A_j[n]$ | The WT approximation coefficient at scale j and location n . |
| f_s | Sampling rate ($f_s = \frac{1}{T_s}$). |
| J | Maximum depth of WT or WPT decomposition. |
| $w_{j,k,n}$ | WT or WPT basis vector at scale j , subband k , and location n . |
| $\Omega_{j,k}$ | The subspace specified by the WT or WPT decomposition at scale j and subband k . |
| \aleph | LDB cost function. |
| $\Gamma_c(j, k, n)$ | The average energy of the WPT coefficients at scale j , subband k , and location n for all patterns in class c . |
| O | STFT window length (samples). |
| Δt | STFT window overlap (samples). |
| | Time resolution. |

| | |
|-------------------------------|--|
| Δf | Frequency resolution. |
| g, h | The highpass and lowpass WT or WPT quadrature mirror filters. |
| $\mathfrak{I}, \mathfrak{J}'$ | Training set, test set. |
| ANN | Artificial Neural Network. |
| AO | The Add-On method of feature selection. |
| BCM | The Bienenstock-Cooper-Munro neuron used to model visual plasticity. |
| CPT | Cosine Packet Transform. |
| CS | The Class Separability method of feature selection. |
| CWT | Continuous Wavelet Transform. |
| DPSS | Discrete prolate spheroidal sequences (Slepian sequences). |
| DWT | Discrete Wavelet Transform (synonymous with WT in this work). |
| KO | The Knock-Out method of feature selection. |
| LDA | Linear Discriminant Analysis. |
| LDB | Local Discriminant Basis. |
| LE | The Local Extrema method of WT dimensionality reduction. |
| MAV | Mean Absolute Value (time domain feature). |
| MAVS | Mean Absolute Value Slope (time domain feature). |
| MES | Myoelectric Signal. |
| MLP | Multilayer Perceptron. |
| MUAP | Motor Unit Action Potential. |
| MVC | Maximum Voluntary Contraction. |
| PCA | Principal Components Analysis. |
| PP | Projection Pursuit. |
| QMF | Quadrature mirror filter. |
| SE | The Subband Energy method of WT dimensionality reduction. |
| SSC | Slope Sign Changes (time domain feature). |
| STFT | Short-time Fourier transform. |
| STTT | Short-time Thompson transform. |
| TD | Hudgins' Time Domain feature set. |
| TFR | Time-Frequency Representation. |
| WL | Waveform Length (time domain feature). |
| WPT | Wavelet Packet Transform. |
| WT | (Discrete) Wavelet Transform. |
| WVD | Wigner-Ville Distribution. |
| ZC | Zero Crossings (time domain feature). |

Acknowledgements

The author would like to acknowledge the financial support of the Natural Sciences and Engineering Research Council, the Whitaker Foundation, the O'Brien Foundation and Hugh Steeper Limited.

It is with sincere gratitude and respect that the author acknowledges the support and guidance of his supervisors, Philip Parker and Maryhelen Stevenson. Their role goes far beyond mentorship; indeed, by example they have instilled a strong appreciation of the work ethic, curiosity, creativity and honesty that is essential if one is to enjoy a career as an academic. Special thanks must also go to Bernie Hudgins who, although not officially a supervisor, was a tremendous influence with respect to learning, exploration, and enthusiasm.

The author is grateful for the support and friendship of the students, staff and faculty of the Institute of Biomedical Engineering and the department of Electrical and Computer Engineering.

The author must take this opportunity to express his deepest gratitude and respect to his parents, Joan and Maurice Englehart. Thank you for making me believe in myself, and not letting me stray far from the tree.

This work simply could never have been completed without the love, support and understanding of my wife, Janet. The time and distraction imposed by a thesis affects those who love you as much as yourself, and Janet, for your patience, I am forever grateful.

To Janet and Benjamin.

Chapter 1

Introduction

1.1 Objectives

The primary objective of this work is to improve the accuracy with which transient patterns of myoelectric signal (MES) activity may be classified, and to do so in a computationally efficient manner. The direct motivation of this research is to enhance a myoelectric control scheme based upon classification of these patterns. Central to this work is the assertion that signal representation plays a crucial role in classification performance. Specifically, it is proposed that time-frequency representations may offer this improved performance.

In an effort to provide improved performance, this work seeks to generalize the transient MES classification problem. Correspondingly, some secondary objectives result:

- Demonstrate the relative efficacy of different forms of signal representation, explain the factors that affect their performance, and specify their optimum configuration for transient MES classification.

- Establish a better understanding of the structural properties of the transient MES. Specifically, explain the intra-class variance that characterizes the transient MES problem, and the demands that this places upon the signal representation.
- Provide an assessment of what the performance bounds might be for the transient MES classification problem, and compare the performance of the methods proposed within to these bounds.

This work accomplishes these objectives using a complement of empirical and theoretical investigation.

1.2 Myoelectric Control

The myoelectric signal, collected at the skin surface, has become an important tool in rehabilitation due to the ease with which it may be acquired. The MES provides information about the neuromuscular activity from which it originates, and this has been fundamental to its use in clinical diagnosis, and as a source of control for assistive devices and schemes of functional electrical stimulation. The signal is utterly complex however, as it is influenced by many factors due to the electrophysiology and the recording environment.

It is the complexity of the MES that has presented the greatest challenge in its application to the control of powered prosthetic limbs. Whether or not an artificial limb is an acceptable replacement for a human limb depends upon the expectations of the affected individual, their motivation to incorporate this device into their lifestyle, and the functionality of the device. Clearly, these issues are not mutually exclusive: a potential user will be more motivated to learn how to use a highly functional device.

Many myoelectric control systems are currently available that are capable of controlling a single device in a prosthetic limb, such as a hand, an elbow, or a wrist. These systems extract control information from the MES based on an estimate of the amplitude [Dorcas66] or the rate of change [Childress69] of the MES. This information is used to specify the function to be performed: the *state* of the device. Once the state is selected, it may be driven at a constant speed, or its speed may be controlled in a manner proportional to the myoelectric activity [Parker86]. These systems have been successful because they require relatively little control information. The extension to controlling multiple functions has

been a much more challenging problem however. Amplitude and rate coded schemes do not provide sufficient information to reliably control more than one or two devices [Vodovnik67]. Unfortunately, multifunction control is a requirement of those with high-level amputations, and these are the individuals who could stand to benefit the most from a functional replacement of their absent limbs.

There have been many attempts to increase the number of states available from the surface MES. Some schemes have used many sites (or *channels*) of amplitude coded information [Schmeidl77][Wirta78][Almstrom81] or other statistical measures [Saridis82]. A vector of features may then be subject to some form of pattern recognition to assign the state. The requirement of several electrode sites, however, introduces severe problems in locating and maintaining the integrity of patterns of MES activity. Others have attempted to increase the capacity of information from one or two channels of MES activity by using time-series models [Graupe82][Doershuk83]. The results were promising, but the method was sensitive to changes in signal amplitude.

In each of these cases, the *steady-state* MES (that produced during constant effort) was used. The steady-state MES however, has very little temporal structure due to the active modification of recruitment and firing patterns needed to sustain a contraction [DeLuca79]. This is due to the establishment of feedback paths, both intrinsic (the afferent neuromuscular pathways) and extrinsic (the visual system). In a departure from conventional steady-state analysis, Hudgins [Hudgins91] [Hudgins93] investigated the information content in the *transient* burst of myoelectric activity accompanying the onset of sudden muscular effort. It was found that significant temporal structure exists in these transient MES bursts and that this temporal structure encodes information important for pattern

discrimination. Hudgins devised a control system for powered upper-limb prostheses using time-domain features (zero crossings, mean absolute value, mean absolute value slope and trace length) and a simple multilayer perceptron (MLP) artificial neural network as a classifier. This controller identified four types of muscular contraction using signals measured from the biceps and triceps. A block diagram of Hudgins' control scheme is shown in Figure 1.1.

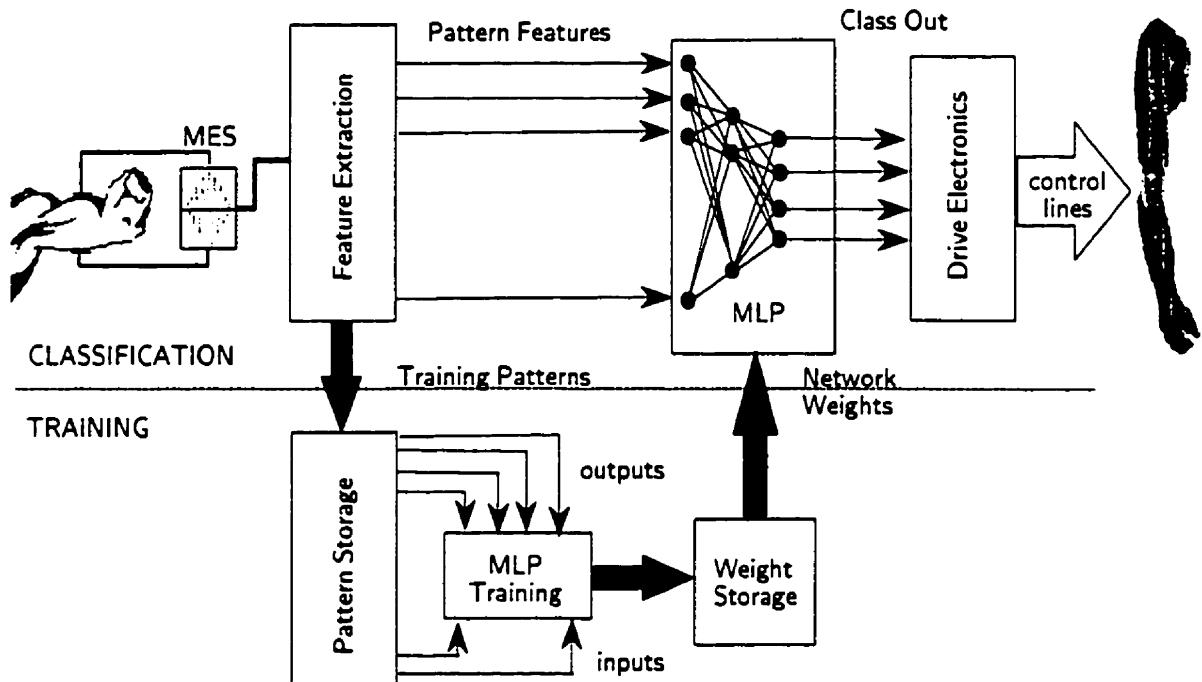


Figure 1.1 – Hudgins' multifunction control scheme. A set of time-domain features is extracted from a transient burst of one-channel MES. A MLP classifier is trained upon an ensemble of patterns derived from contractions of (up to) four movement types.

Not only does this system provide multifunction control from a single site, but the control signals can be derived from natural contractions, thereby minimizing the conscious effort of the user. This system performs well¹, but improved classification performance would benefit the functionality and ultimately, the acceptance of artificial limbs controlled by the MES.

¹ Hudgins demonstrated an average classification performance of 89% on an ensemble of 15 subjects, including nine normally-limbed subjects and six limb-deficient individuals.

In the quest to improve classification accuracy, one has the choice of improving the classifier or the means of signal representation (the feature set). Although some classifiers demonstrate obvious advantages over others, it is the signal representation that most dramatically affects the classification performance [Fukunaga90][Bishop96], and this is the focus here. Given that transient MES patterns have structure in both time and frequency, it is suggested that the information which would best discriminate amongst contraction types would be concentrated in a dual representation. Consequently, this work explores the efficacy of feature sets derived from time-frequency representations.

1.3 The Transient MES

Fundamental to the successful classification of the transient MES is an understanding of its origin. The following sections describe the physiological processes from which these patterns arise.

1.3.1 The Etiology of the MES

The functional unit of a skeletal muscle is the *motor unit*, which consists of a group of muscle fibres and the terminal branches of a common nerve axon that innervates them [DeLuca79]. The nerve axon has a cell body (the *motor neuron*) in the anterior horn of the spinal cord. This is illustrated in Figure 1.2.

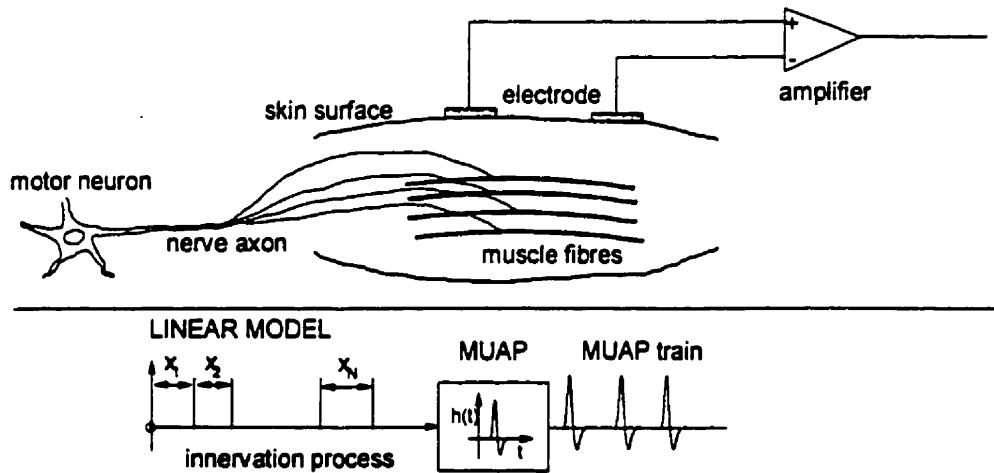


Figure 1.2 – A single motor unit. A motorneuron innervates a group of muscle fibres, constituting a motor unit. A linear model is shown, in which a series of excitatory impulses (separated by the intervals x_1, x_2, \dots) from a motor neuron innervate a group of muscle fibres. The motor unit action potential (MUAP) is the electrical response of each impulse, and the series generates a MUAP train.

In voluntary contractions, a release of acetylcholine at the nerve-muscle synapse (the endplate) initiates a fundamental biochemical process that actuates a

contractile twitch of the muscle fibre. Congruently, a depolarization of the muscle fiber membrane generates an electromagnetic field that propagates through the conducting tissue. An electrode placed in proximity to the motor unit will record the summation of the response from the innervated fibres. This pulse is referred to as the *motor unit action potential* (MUAP).

The duration of each action potential is inversely related to the conduction velocity of the muscle fibres, which can range from 2 to 6 m/s, in inverse relation to the fibre diameter [Buchthal54]. The propagation of the action potential through the surrounding muscle tissue has a low-pass filtering effect, commonly referred to as a *tissue filter* [Lindstrom77]. Therefore, the primary factors that determine the shape of a MUAP are the diameter and geometrical arrangement of the constituent muscle fibres, the tissue filtering effect, and the properties of the recording electrode and instrumentation.

In order to sustain a voluntary muscle contraction, the motor units must be repeatedly activated. This innervation process evokes a series of MUAPs, or a MUAP *train*. Unless force levels are very low, and/or the electrode is very selective, it is likely that multiple motor units will influence a recording. The surface MES generally comprises the temporal and spatial superposition of many MUAP trains.

1.3.2 The MES During Sustained Contractions

Control of skeletal muscle force is achieved primarily by [Bigland54][Milner73]:

- i) *recruitment*: varying the number and composition of activated motor units, and
- ii) *firing rate*: varying the rate of activation of the individual motor units.

During force-varying contractions, recruitment is the dominant factor at the beginning of a contraction, with the smallest motor units being recruited first (*the size principle* [Henneman65]). A motor unit at the threshold of recruitment has an unstable firing rate, with a minimum value of roughly 5 pulses/s [Person72]. At up to 30% of maximum voluntary contraction (MVC), recruitment remains the dominant factor, with larger motor units recruited as force increases. Between 30% and 75% MVC, the firing rate changes from a secondary to a primary role in force production. At above 75% MVC, further force increases rely almost solely on firing rate increments, with negligible recruitment activity [DeLuca72].

The complexity of the MES is due to its origin: it is the composite of many sources, each of which is affected by many factors. During a sustained contraction, the interaction between recruitment and firing rate yields a signal that contains essentially no structure in its instantaneous temporal waveform [Parker86]. This limits the amount of control information that can be derived from a steady-state signal.

1.3.3 Structure in the Transient MES

While investigating the properties of the MES coincident with the onset of rapid contractions, Hudgins [Hudgins91] observed a substantial degree of structure in the transient waveforms. Data were acquired during small but distinct isometric and anisometric contractions, using a single bipolar electrode pair placed over the biceps and triceps muscle groups. This arrangement was intended to allow a large volume of musculature to influence the measured activity. Figure 1.3 shows typical patterns corresponding to flexion/extension of the elbow, and pronation/supination of the forearm.

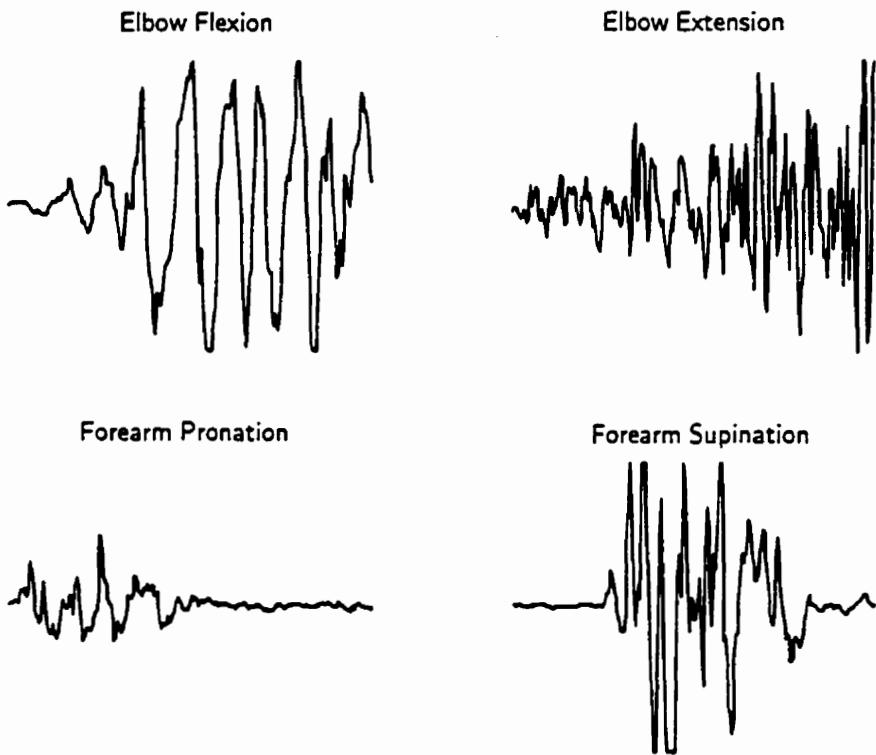


Figure 1.3 – Patterns of transient MES activity recorded using a single bipolar electrode pair, placed over the biceps and triceps.

These patterns exhibit distinct differences in their temporal waveforms. Within a set of patterns derived from the same contraction, the structure that characterizes the patterns is sufficiently consistent to maintain a visual distinction between

different types of contraction. Hudgins aligned the patterns using a cross-correlation technique and showed that the ensemble average of patterns within a class preserves this structure.

The question is: why do these structured patterns occur? The amplitude envelopes of these classes of contraction certainly differ, but the degree of structure in the patterns cannot be explained by the gross activity of different muscle groups. This structure suggests that it is likely that a “motor plan” exists for simple, ballistic contractions. In this scenario, an orderly scheme of recruitment and neural discharge is responsible for initiating a contraction. The absence of concrete evidence of this phenomenon is due to the difficulty in identifying motor unit recruitment and firing activity in such a short, dynamic interval. We may make the following observations, however:

1. Weirzbicka *et al.* [Weirzbicka93] have identified that the onset of rapid contraction (both isometric and anisometric) produces a transition from tonic to phasic² muscle activation, and that a brief interval of silence (inactivity) exists between these states. This suggests the initiation of a new motor program.
2. It has been observed that motor unit recruitment order appears stable for a given task, once the task has been learned [Basmajian85]. Further, consistent feedforward commands from the motor cortex have been observed during goal-directed movements, such as arm positioning tasks [Georgopoulos93].

² Tonic motor units are those responsible for *sustaining* force; phasic motor units are those responsible for *building* force.

3. The propagation delays in the neuromuscular system's afferent feedback pathways prohibit modification of motor activity immediately following initiation of a spontaneous contraction. The *stretch reflex* regulates motion as negative-feedback servomechanism, relying on proprioceptors which sense muscle length and tension. The *short-loop* stretch reflex completes its feedback path at the spinal level, resulting in a propagation delay of 30-50 ms. The *long-loop* stretch reflex must consult the cerebellum, resulting in a delay of 50-80 ms [Schmidt88]. Therefore, the neuromuscular system is operating either partially or fully in an "open-loop" condition for the first 30-80 ms [Atsma97]. Indeed, the greatest degree of structure is seen in this early portion of the transient MES patterns.

It is possible that the structure may be due to factors that are not of neuromuscular origin. An obvious mechanism that could impose a deterministic effect is movement artifact (motion of the electrodes relative to the skin). Attempts to reproduce the patterns by passive movement of the limb and rapid electrode lead motion however, have failed [Hudgins93]. Yamazaki *et al.* [Yamazaki93] simultaneously recorded patterns of rapid isometric contraction of the biceps using bipolar surface electrodes and needle electrodes inserted into the muscle. The needle electrode recordings were highly correlated with the surface recordings, demonstrating fairly conclusively that the activity is not due to movement artifact. Other possible sources of modulation, such as the motion of the active muscle fibers (relative to the electrode) have not been disproved.

Subtle changes in the nature of a contraction however, can introduce variability into the recorded MES. In an ensemble of patterns produced by similar contractions, there are visually perceptible similarities amongst waveforms, but

the local characteristics may vary tremendously. Identifying this loosely defined structure is a challenging pattern recognition task.

1.3.4 Two Channel Transient MES Patterns

It was shown by Kuruganti *et al.* [Kuruganti95] that the performance of Hudgins' myoelectric control system could be enhanced by using two channels of localized MES activity instead of one channel of global activity. This implies that the primary sources of information are the activity of the biceps and triceps, and that some information is lost by destructive superposition of activity in Hudgins' widely-spaced electrode arrangement. The localized activity of the biceps and triceps, collected using two sets of closely spaced bipolar electrode pairs, is shown in Figure 1.4.

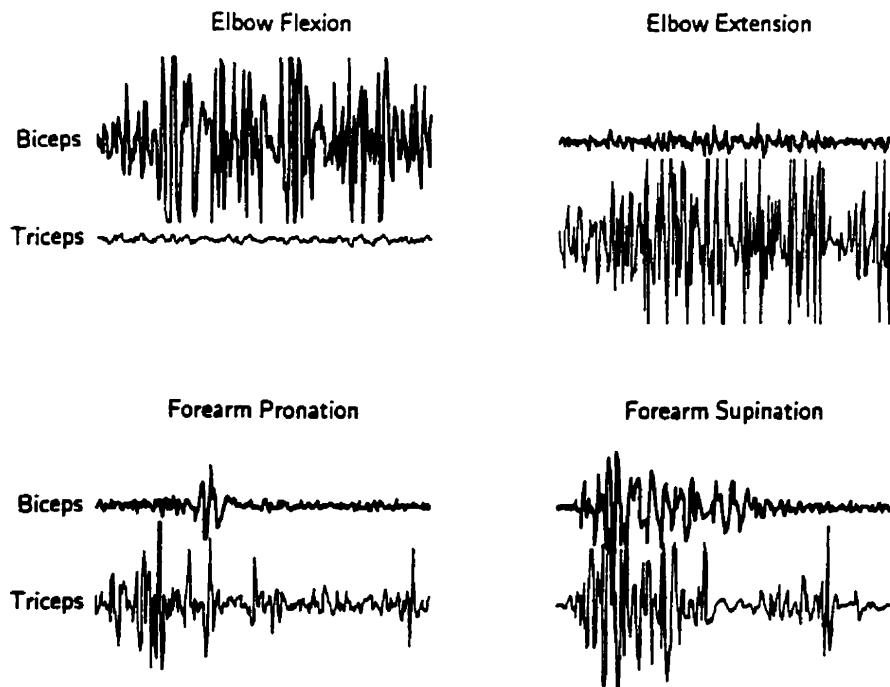


Figure 1.4 – Patterns of transient MES activity recorded using two sets of closely spaced bipolar electrode pairs, placed over the biceps and triceps.

Clearly, localizing the activity of the biceps and triceps provides a greater distinction between the classes than a single channel with the superimposed activity. In the interest of providing the most informative representation of the transient MES, it is this two channel configuration that will be used in this work.

1.4 Thesis Outline

Chapter 1 has provided an introduction to the problem of myoelectric control, emphasizing that the primary factor limiting dexterity is the amount of information that can be derived from the MES. A brief background of the etiology of the MES has been given, with the intention of postulating that the structure inherent in the transient MES may be due to some determinism in its neuromuscular origin. Previous work has demonstrated that this structure may be exploited to yield good classification performance [Hudgins91]. It is proposed in this work that this structure is manifest in the time-frequency domain, and that time-frequency representations may improve the classification performance.

Chapter 2 presents detailed discussion signal representation for pattern classification. The problem is broken down into the tasks of *feature extraction*, *dimensionality reduction*, and *classification*. The stages of feature extraction and dimensionality reduction comprise the procedure of *signal representation*. It is emphasized that, although an appropriate classifier is necessary, it is the signal representation that profoundly affects the classification performance of a given problem. Feature sets based upon the short-time Fourier transform, the wavelet

transform, and the wavelet packet transform are introduced as candidates for real-time implementation. The importance of dimensionality reduction is introduced, and alternative methods are characterized as belonging to one of two methodologies: *feature selection* or *feature projection*. The relative performance of these two approaches provides important insight throughout the thesis.

Chapter 3 provides a mathematical background of wavelet and wavelet packet transforms. In addition to basic theory, the means of deriving feature sets for classification is described.

Chapter 4 presents the classification results of time-frequency based feature sets. The focus of these analyses center upon a database of two channel transient MES patterns acquired from 16 subjects. For each feature set, the optimal transform parameters and dimensionality reduction strategies are determined empirically. The conclusion of these analyses is a specification of the best signal representation for the transient MES problem.

Chapter 5 provides some perspective on the results of Chapter 4. The theoretical performance bounds for the transient MES classification problem are described, and the performance of the time-frequency based signal representations are compared to this bound. This chapter also provides a model that seeks to explain the loose structure of the transient MES.

The observations and conclusions of this research are summarized in Chapter 6. From these, the original contributions are accounted. The concluding discussion includes recommendations for future investigation which may build upon this work.

Chapter 2

Signal Representation for Classification

2.1 Introduction

Signal classification, signal compression and noise removal are examples of classic signal analysis problems. Each has been widely studied, and each is rich with theory and applications. Of vital importance to any aspect of signal analysis, however, is the means by which the signal(s) are represented. In this chapter, we investigate the means by which relevant features may be extracted and irrelevant information discarded, in the context of signal classification.

Often, important features for classification are characterized by local information in the dual domains of time and frequency. This is especially true for transient signals, such as the MES patterns under investigation here. It is demonstrated here that the time-frequency domain provides an effective setting for constructing feature sets for pattern classification. Moreover, fast algorithms exist that allow

efficient computation of time-frequency methods, allowing *real-time* feature extraction for MES classification.

The organization of this chapter is as follows. Section 2.2 provides a formulation of the problem of *feature extraction* and *pattern classification*. Section 2.3 introduces the reader to pattern classification, and describes the advantages and drawbacks of established techniques. In Section 2.4, the importance of feature extraction to the success of pattern classification is discussed in detail. Section 2.5 focuses upon *time-frequency representations*, and the issues relevant to their application as feature extractors. *Dimensionality reduction* is often a necessary complement to feature extraction, especially when using time-frequency methods. Dimensionality reduction techniques tailored for use in classification problems are discussed in Section 2.6. Finally, Section 2.7 summarizes the chapter, and proposes that time-frequency methods, coupled with appropriate dimensionality reduction, constitute a robust signal representation for classification of the transient MES. This proposition is the focal point of this thesis.

2.2 Problem Formulation

It is useful to establish some notation to describe the classification problem. A *pattern* may be said to consist of N variables $\{x_i\}_{i=1}^N$. It is convenient to gather the variables x_i together and denote them by a single vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_N]^T$. This is the *measurement vector* which, in its simplest form, may be the elements of a

sampled signal¹. Each pattern \mathbf{x} may be said to belong to one of K classes, denoted as y_k . We may say then, that $\mathbf{x} \in \mathcal{X} \subseteq \Re^N$ is the *input signal space* and $y \in \mathcal{Y} = \{y_1, \dots, y_K\}$ is the *output response space*, which is simply a collection of K class labels. Therefore, $\mathcal{X} \times \mathcal{Y}$ is the set of all pairs of *input signals* and the corresponding *class labels* (\mathbf{x}, y) . Signal classification may be regarded as a function $d: \mathcal{X} \rightarrow \mathcal{Y}$, which assigns a class label to each input signal $\mathbf{x} \in \mathcal{X}$.

Direct application of the data in signal space is usually prohibitive because (*i*) the signal space often has a high dimension (the MES patterns here nominally have a dimension of $N = 256$), and (*ii*) the presence of noise (or other unwanted components) makes classification difficult. Indeed, the signal space is highly redundant with respect to the response space. This implies the need to reduce the dimensionality of the problem; one must extract only the *features* needed to discriminate the signals, and discard everything else. This can greatly improve the performance of a chosen classifier, and reduce its complexity. To this end, we define a *feature space* $\mathcal{F} \subset \Re^M$ between the signal space and the response space, where $M \leq N$. A *feature extractor* is defined as a map $f: \mathcal{X} \rightarrow \mathcal{F}$, and the classifier as a map $g: \mathcal{F} \rightarrow \mathcal{Y}$. The classification process may then be denoted as $d = g \circ f$.

In general, we assume the presence of a *training dataset* \mathfrak{I} , which consists of P pairs of input signals and class labels (\mathbf{x}, y) :

¹ Conventional notation in mathematics has a measured signal denoted as a vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_N]^T$. Signal processing literature tends to represent this measurement as a time series: $\mathbf{x} = \{x[1], \dots, x[N]\}$. These two forms will be used interchangeably, taking the form most suited to the context. It is hoped that this will simplify, rather than confuse the notation.

$$\mathfrak{I} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \dots, \left(\mathbf{x}^{(P)}, y^{(P)} \right) \right\}, \quad (2.1)$$

where the superscript $^{(i)}$ denotes the i^{th} pattern-class pair within the training set. Unless required for specificity, this superscript will be omitted.

For the purpose of evaluating the generalization ability of a classifier, we presume the availability of a *test dataset* $\tilde{\mathfrak{I}}$, which may be assumed independent of \mathfrak{I} , but derived from the same probability model.

2.3 Pattern Classification

The task of categorizing observed or measured data into classes is central to many applications. The act of classification is tightly bound to the proper extraction of relevant features from the unprocessed data, as depicted in Figure 2.1:

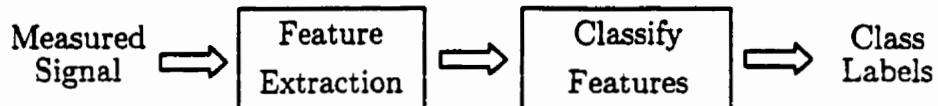


Figure 2.1 – The Classification Problem.

In this section, a brief overview is given of pattern classifiers that have demonstrated theoretical rigor and practical utility. These classifiers are best interpreted in the context of Section 2.4, which is dedicated to the important preliminary stage of feature extraction. It will be shown that feature extraction is fundamental to classifier performance; even the most adept classifier must have an appropriate and efficient representation of the physical process which it must interpret.

The practical methodologies that exist for pattern classification may be loosely grouped into three categories. Historically, the two classical methods are the *statistical* (or *decision-theoretic*) approach [Tou74][Duda73] and the *structural* (or *syntactic*) approach [Fu82]. The third, and most recently established type of pattern classifier is the *learning* (or *neural*) approach. Learning algorithms have their origins in perceptrons [Rosenblatt58] and adaptive linear elements [Widrow90], and have matured into the diverse field of neural networks [Hush93].

Statistical pattern recognition is based upon a statistical analysis of the data to be classified. The data are assigned to a particular class by compiling a probabilistic model (estimating probability density functions) of the data in N -dimensional space, and dividing the space into regions corresponding to each class, according to some criterion. The major accomplishments in statistical pattern recognition include Bayesian classifiers, distance classifiers, and classification and regression trees [Breiman84]. These will be examined in this chapter.

The syntactic approach, on the other hand, is based on utilizing the structure of patterns and the interrelationships between the components of a pattern. Syntactic pattern recognition involves identifying meaningful components or “*primitives*” of the patterns, and developing a formal syntax or “*grammar*” describing the synthesis of the patterns from their primitives. The preference here is to discuss structural methods in the context of feature extraction. From the perspective of this work, the development of primitives and syntax is more a signal representation issue than a classification task. Section 2.4 demonstrates the importance of structural representations in feature extraction.

Learning algorithms, today, almost invariably take the form of artificial neural networks. Artificial neural network approaches may also be termed *deterministic* as opposed to *statistical* because the learning algorithms assume nothing about the statistical properties of the pattern classes. It will be shown, however, that statistical and neural network pattern classifiers are very similar in form and objective.

The emphasis of this work is upon signal representation, not upon classifiers. The intent of this section is to illustrate the major features of the most popular pattern

classifiers in use today, and the differences between them that are important. Correspondingly, two representative classifiers will be chosen to carry through the thesis based upon the ease with which they may be interpreted and their applicability to the MES classification problem.

2.3.1 Bayesian Pattern Classification

The central problem in statistical pattern recognition is the development of decision functions from sets of finite sample patterns of different classes so that the functions will partition the input space into regions, each of which contains the sample patterns belonging to each class. Recall the definition of the classification problem given in Section 2.1: the input signal space is $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^N$ and the response space is $y \in \mathcal{Y} = \{y_1, \dots, y_K\}$, where \mathbf{x} and y are the input pattern and the class label, respectively. The measurements \mathbf{x} and y may be considered in a probabilistic framework, and viewed as single observations of the random variables X and Y . In general, the most information that can be known about the input space are the *a posteriori* probabilities

$$P(y_k | \mathbf{x}) \text{ for } k = 1, \dots, K. \quad (2.2)$$

This is the probability that pattern \mathbf{x} comes from class y_k .

In this framework, pattern classification is posed as a statistical decision problem; one evaluates the K *a posteriori* probabilities and selects the largest. In general, the *a posteriori* probabilities $P(y_k | \mathbf{x})$ are not known, but may be calculated from

the *a priori* probabilities $P(y_k)$ and the conditional density functions $p(\mathbf{x}|y_k)$ using Bayes' theorem, which is [Papoulis85]:

$$P(\mathbf{x}, y_k) = P(y_k)p(\mathbf{x} | y_k) = p(\mathbf{x})P(y_k | \mathbf{x}). \quad (2.3)$$

Rearranging, we get

$$P(y_k | \mathbf{x}) = \frac{P(y_k)p(\mathbf{x} | y_k)}{p(\mathbf{x})} \quad (2.4)$$

where

$$p(\mathbf{x}) = \sum_{j=1}^K P(y_j)p(\mathbf{x} | y_j). \quad (2.5)$$

Note that $p(\mathbf{x})$ is the probability density function of the input space; this remains constant for all $P(y_k | \mathbf{x})$, so it can be ignored for purposes of discrimination. When the true class distributions are not known, the *a priori* probabilities are often made equal: $P(y_k) = 1/K$ for $k = 1, \dots, K$.

To summarize, Bayes' decision rule is really nothing more than the implementation of the decision functions:

$$d_k(\mathbf{x}) = p(\mathbf{x} | y_k)P(y_k), \quad k = 1, \dots, K \quad (2.6)$$

where a pattern \mathbf{x} is assigned to class y_i if for that pattern $d_i(\mathbf{x}) > d_j(\mathbf{x})$ for all $j \neq i$. This Bayes' decision rule has the property that *the probability of classification error is minimized*, making Bayes' classifier statistically superior to any other. The Bayes classifier is illustrated in Figure 2.2.

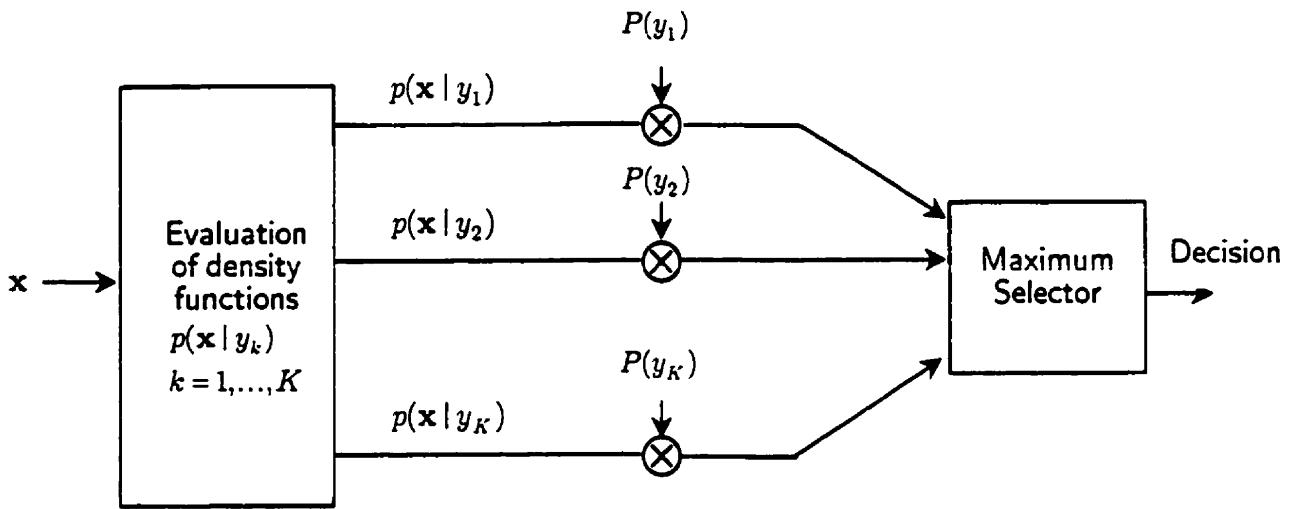


Figure 2.2 – A Bayesian Classifier

The challenge here lies in the estimating the densities $p(\mathbf{x}|y_k)$ from the training data. This is difficult, if not impossible, if the dimension of the input space N is large. The *curse of dimensionality* imposes the constraint that the number of training set examples must be much greater than N to get a reliable estimate of $p(\mathbf{x}|y_k)$.

Comment 2.1. The Curse of Dimensionality

The fact that most physical signals occupying N -dimensional space have an *intrinsic dimension* that is much less than N leads to a paradox. If one attempts to represent multi-class data in a high-dimensional space, an extremely large amount of sample data is necessary to populate the space to adequately define the class boundaries. Indeed, if one attempts to increase the information content of a representation by increasing the dimension, the class neighbourhoods become more sparsely populated, and thus, their boundaries less likely to represent the true class boundaries. Moreover, high-dimensional feature sets are susceptible to noise, and the presence of irrelevant feature combinations can actually obscure the impact of the more discriminating inputs. This has been termed the *curse of dimensionality* [Bellman61].

If some assumptions about the nature of the conditional probabilities can be made, then the estimation of $p(\mathbf{x}|y_k)$ is simplified. An important form of an approximate Bayesian classifier is described in the next section.

2.3.1.1 The Gaussian Bayes Classifier

If it is reasonable to assume a parametric form of the conditional probability density functions $p(\mathbf{x}|y_k)$, then the Bayes classifier derived in the preceding section can take a more tractable form. A common assumption is that the densities $p(\mathbf{x}|y_k)$ are multivariate normal (Gaussian). Although for some datasets it is difficult to make this assumption, the normal distribution does represent an appropriate model for many practical applications.

Consider K classes of patterns, governed by the multivariate normal density functions:

$$p(\mathbf{x} | y_k) = \frac{1}{(2\pi)^{N/2} |\mathbf{C}_k|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mathbf{m}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \mathbf{m}_k)\right], \quad k = 1, \dots, K \quad (2.7)$$

where each density is completely specified by its mean vector \mathbf{m}_k and its covariance matrix \mathbf{C}_k , which are defined as

$$\mathbf{m}_k = E_k[\mathbf{x}] \quad (2.8)$$

and

$$\mathbf{C}_k = E[(\mathbf{x} - \mathbf{m}_k)(\mathbf{x} - \mathbf{m}_k)^T] \quad (2.9)$$

where $E_k[\cdot]$ denotes the expectation operator over the patterns of class y_k . Here, $|\mathbf{C}_k|$ indicates the determinant of matrix \mathbf{C}_k . Sample patterns taken from a normal distribution tend to fall in a single cluster with its center determined by the mean

vector and its shape defined by the covariance matrix. The loci of points of constant density are hyperellipsoids with the principal axes in the directions of the eigenvectors of the covariance matrix and the lengths of these axes determined by the eigenvalues.

According to Equation (2.6), the decision function for class y_k may be chosen as $d_k(\mathbf{x}) = p(\mathbf{x} | y_k)P(y_k)$. Because of the exponential nature of the normal density function however, it is more convenient to work with the natural logarithm of this decision function. In other words, we may use the form

$$d_k(\mathbf{x}) = \ln[p(\mathbf{x} | y_k) \cdot P(y_k)] = \ln p(\mathbf{x} | y_k) + \ln P(y_k). \quad (2.10)$$

Substituting Equation (2.7) into Equation (2.10) yields

$$d_k(\mathbf{x}) = \ln P(y_k) - \frac{N}{2} \ln 2\pi - \frac{1}{2} \ln |\mathbf{C}_k| - \frac{1}{2} [(\mathbf{x} - \mathbf{m}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \mathbf{m}_k)], \quad k = 1, \dots, K \quad (2.11)$$

Since the term $\frac{N}{2} \ln 2\pi$ does not depend on k , it can be eliminated, giving

$$d_k(\mathbf{x}) = \ln P(y_k) - \frac{1}{2} \ln |\mathbf{C}_k| - \frac{1}{2} [(\mathbf{x} - \mathbf{m}_k)^T \mathbf{C}_k^{-1} (\mathbf{x} - \mathbf{m}_k)], \quad k = 1, \dots, K \quad (2.12)$$

which is Bayes' decision function for normally distributed patterns. These decision functions are *hyperquadratic* since the last term contains a second order of \mathbf{x} , meaning that the best that this Bayesian classifier can do is to place a *quadratic discriminant function* between pattern classes. If the pattern classes are truly characterized by normal densities, however, no other surfaces will yield better results, on an average basis. The quadratic decision functions are

$$d_k(\mathbf{x}) = \ln P(y_k) - \frac{1}{2} \ln |\mathbf{C}_k| - \frac{1}{2} \mathbf{x}^T \mathbf{C}_k^{-1} \mathbf{x} + \mathbf{x}^T \mathbf{C}_k^{-1} \mathbf{m}_k - \frac{1}{2} \mathbf{m}_k^T \mathbf{C}_k^{-1} \mathbf{m}_k. \quad k = 1, \dots, K \quad (2.13)$$

If we assume a pooled covariance matrix², \mathbf{C} , it follows that the decision functions become

$$d_k(\mathbf{x}) = \ln P(y_k) + \mathbf{x}^T \mathbf{C}^{-1} \mathbf{m}_k - \frac{1}{2} \mathbf{m}_k^T \mathbf{C}^{-1} \mathbf{m}_k \quad (2.14)$$

$$k = 1, \dots, K$$

which represents a set of linear discriminant functions. In this case, the decision surface is linear with respect to the input space, describing a hyperplane. The normal Bayesian classifier is therefore often called a *linear discriminant analysis* (LDA). It can be shown that linear and quadratic discriminant functions derived from this Bayesian classifier are theoretically optimal for distributions other than those that are Gaussian [Tou74].

Some constraints relevant to the applicability of a LDA are:

1. The assumptions of normal pattern densities must be reasonable.
2. The data must be reasonably well clustered and linearly separable.
3. It is sensitive to outliers and noise, which makes it vulnerable to the curse of dimensionality.

The advantages of the LDA are:

1. It is easily interpreted and implemented.
2. It trains very quickly with reasonably-sized datasets.
3. It requires no adjustment of its architecture or training algorithm.

These advantages make it an attractive choice for clinical implementation.

² In this work, the pooled covariance matrix has been computed as $\mathbf{C} = \frac{1}{K} \sum_{k=1}^K \mathbf{C}_k$.

2.3.2 Artificial Neural Networks

An artificial neural network (ANN) is a computational system inspired by the learning characteristics and the structure of biological neural networks. From humble beginnings about 50 years ago, neural network theory has matured to the point that the ANN has become an irreplaceable tool in many applications and disciplines. The application of ANNs as pattern classifiers is described in this section. Figure 2.3 shows an hierarchy of some artificial networks that have been used as pattern classifiers.

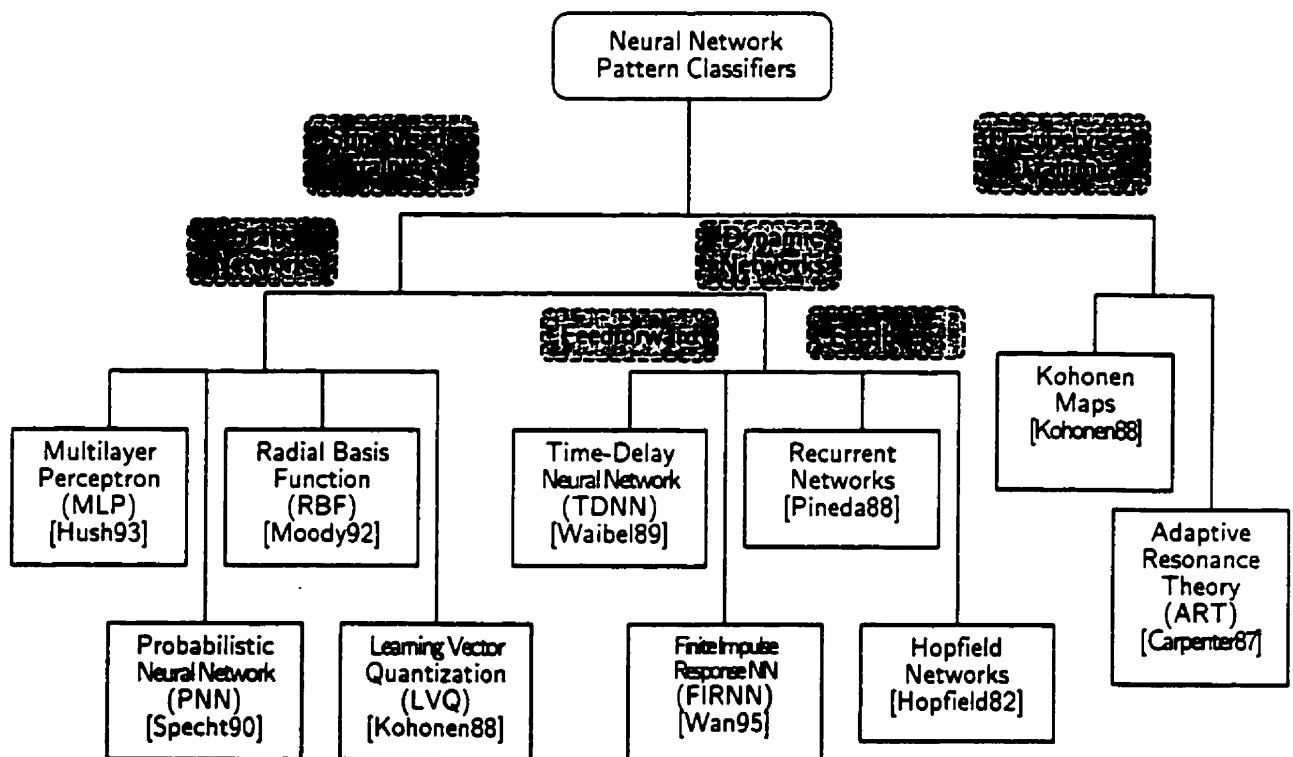


Figure 2.3 – An hierarchy of some artificial neural networks that have been used as pattern classifiers.

A discussion of individual networks will not be given here; reference for more detail is given in the figure.

For our purposes here, the discussion of ANNs will be limited to those trained using *supervised learning*. This means that they are presented with a training set of M example pairs from the input space and the response space:

$$\mathfrak{I} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(P)}, y^{(P)})\}, \quad (2.15)$$

where P is finite. If class membership information is available during training, supervised methods, in general, will fare better than unsupervised methods. This is due to the fact that knowledge of class membership aids the construction of appropriate discriminant boundaries. The ability of the ANN to handle independent datasets is assessed by providing a test set \mathfrak{I}' , which is presumed to be a reasonable representative of data the network will see in actual use.

The strength of neural network based pattern classifiers lies in their applicability to problems involving arbitrary distributions of data. Moreover, a firm understanding of the pattern recognition properties of neural networks has emerged, relating their characteristics to Bayesian decision making [Makhoul91]. ANN classifiers often yield classification rates comparable to Bayesian methods without *a priori* information.

Of all ANN architectures that have been used as pattern classifiers, the most commonly used is the multilayer perceptron (MLP). In turn, the learning algorithm which is almost always used to train the MLP is the backpropagation algorithm, which is a stochastic approximation of the steepest descent algorithm [Haykin94]. The MLP architecture and the backpropagation algorithm are the simplest, and most extensively studied of all neural network paradigms. A MLP containing nonlinear activation functions is capable of constructing arbitrarily complex decision boundaries in feature space for networks of two layers or more. Some problems associated with the MLP and backpropagation are that training

may be slow, and that selection of the best network size may be difficult [Makhoul91].

Neural network pattern classifiers more advanced than the MLP have been shown to offer better training characteristics, greater immunity to noise, better response with small training sets, and an ability to better handle high-dimensional inputs [Hush93]. Specialized architectures and training algorithms however, are not of interest here. Indeed, it has been demonstrated that classification error rates are similar across different classifiers when they are powerful enough to form minimum error decision regions, when they are properly tuned, and when sufficient training data is available [Ghosh92]. For the purpose of assessing the discriminability of MES patterns using a variety of signal representations, a MLP network trained using a standard backpropagation algorithm will suffice. The following sections introduce the MLP, the means by which it is trained, and its capabilities.

2.3.2.1 The Perceptron

The fundamental computational unit is the *perceptron*, conceived by Rosenblatt [Rosenblatt58]. As depicted in Figure 2.4, the perceptron forms a weighted sum of the n components of the input vector $\mathbf{x} = [x_1 \ x_2 \ \dots \ x_N]^T$ and adds a *bias value*, θ . The result is then passed through a nonlinearity $f(\bullet)$.

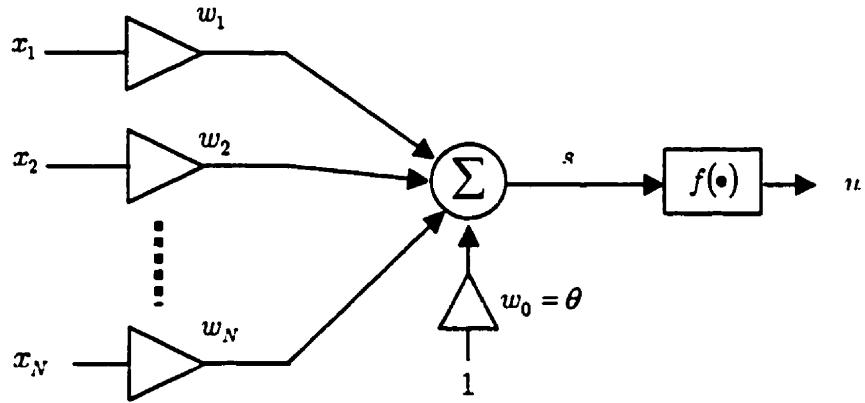


Figure 2.4 – The Perceptron.

Rosenblatt's original model used a *hard-limiting nonlinearity*:

$$f_{HL}(s) = \begin{cases} 1 & s > 0 \\ 0 & s \leq 0 \end{cases}. \quad (2.16)$$

which is illustrated in Figure 2.5. When perceptrons are combined together in layers, it is more common to use the *logistic sigmoid nonlinearity*:

$$f_{\text{log}}(s) = \frac{1}{1 + e^{-\beta s}} \quad (2.17)$$

This function is continuous and varies monotonically from 0 to 1 as s varies from $-\infty$ to ∞ . The gain of the sigmoid, β , determines the steepness of the transition region; this is often set to 1. The main advantage of the sigmoid nonlinearity is that it is *differentiable*. This property has had an historical impact because it made it possible to derive a gradient search algorithm for networks with multiple layers.

Another function belonging to the sigmoid family is the *hyperbolic tangent sigmoid*:

$$f_{\text{tanh}}(s) = \frac{e^{\beta s} - e^{-\beta s}}{e^{\beta s} + e^{-\beta s}}, \quad (2.18)$$

the outputs of which range from -1 to 1 . In many cases, networks which use the hyperbolic tangent sigmoid as a nonlinearity tend to learn faster than those which use the logistic sigmoid [Haykin94].

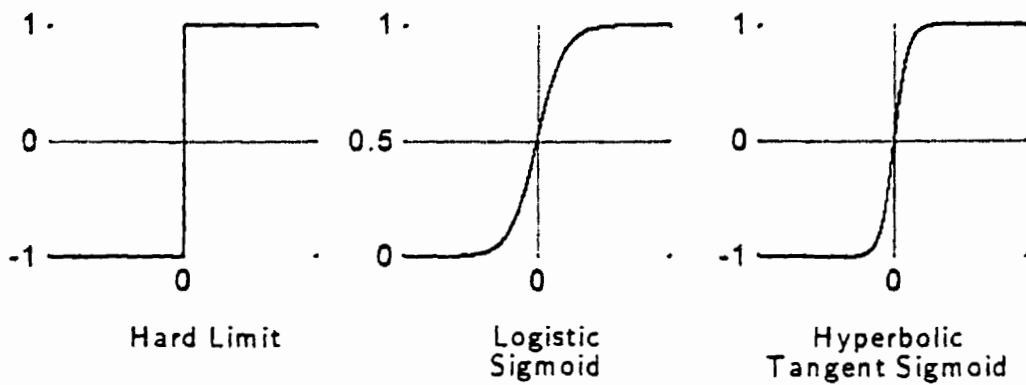


Figure 2.5 – Three common types of nonlinearity used as the activation function in an artificial neuron.

2.3.2.2 The Multilayer Perceptron

The capabilities of single perceptrons however, are limited to linear decision boundaries, and are suitable only for problems requiring a simple linear dichotomization of the pattern space. Many problems require a nonlinear partitioning of the pattern space. This can be achieved using a multilayer perceptron network, which cascades two or more layers of perceptrons together, making it possible to partition the pattern space with arbitrarily complex decision boundaries. The individual perceptrons in the network are called *neurons* or *nodes*, and usually employ a sigmoid nonlinearity instead of a hard limiter. A typical MLP network architecture is depicted in Figure 2.6.

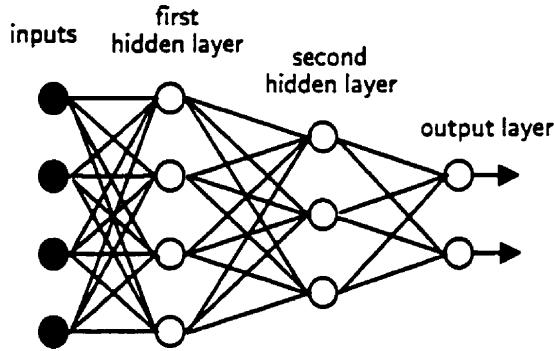


Figure 2.6 – The architecture of a typical MLP network.

The input vector feeds in to the first layer nodes; the outputs of this layer feed into each of the second layer nodes, and so on. Often, the nodes are fully connected between layers. The terminology used here will not include the input vector as a layer, so that the network in the figure above is referred to as a three-layer network. The multiple nodes in the output layer correspond to multiple classes in a pattern recognition problem.

For classification problems, Lippmann [Lippmann87] demonstrated that a 2-layer MLP can implement arbitrary convex decision boundaries, given a sufficient number of hidden layer nodes. Essentially, each hidden layer node provides a linear boundary in pattern space, and each of the boundaries may be nonlinearly connected in a smooth fashion with the others by the sigmoid nonlinearity.

Many algorithms have been developed which adapt the network weights so as to provide a suitable map between the set of input vectors and the set of desired responses. In general, an algorithm is either *supervised*, in which the desired response is available during the learning phase, or *unsupervised*, in which clusters are formed from the input patterns. A training dataset of transient MES patterns includes knowledge of the actual class of movement, and therefore, our interest

here is limited to supervised learning. The backpropagation algorithm will be used to train the MLP in this work. A derivation of the algorithm has been included in Appendix A.

2.3.2.3 Issues in MLP ANN Training

Learning Rate. The learning rates can be uniform throughout the network, or different for each layer or node. In general, it is difficult to determine the best learning rate, but a useful rule of thumb is to make the learning rate for each node inversely proportional to the average magnitude of vector feeding the node. Many schemes which *adapt* the learning rate as a function of the local curvature of the error surface have been proposed [Jacobs88]. The simplest approach, and one that works quite well in practice, is to add a *momentum* term of the form $\alpha(w_{t,j,i}(k) - w_{t,j,i}(k-1))$ to each weight update, where $0 < \alpha < 1$. This term makes the current search direction an exponentially weighted average of past directions, and helps keep the weights moving across flat portions of the error surface after they have descended from steep portions.

Stopping Criteria. The iterative process of computing the gradient and adjusting the weights is continued until a minimum is found in the error surface (or a point determined to be sufficiently close). Several measures are candidates for stopping criteria. If the magnitude of the gradient falls below a chosen level, the algorithm may be terminated, as this may indicate that the minimum is being approached. Perhaps a more common stopping criterion is a lower threshold on the sum squared error, $J(w)$. This requires knowledge (or a decent estimate) of

the minimum value of $J(w)$, which is not always available. One might consider stopping when a chosen number of iterations have been performed. In this situation the number of iterations must be determined by empirical evidence gathered from previous training sessions.

Finally, the method of *cross-validation* can be used to monitor the *generalization* performance during learning. This method entails splitting the data into a training set and a test set. During learning, the performance of the network upon the training set can only improve, but the performance on the test data will only improve to some point, beyond which it will start to degrade. It is at this point that the network starts to *overfit* the data, and the algorithm should be terminated. In pattern recognition applications, the method of cross validation may be extended to monitoring the classification rate of the network upon the training set and the test set data, stopping the algorithm when the classification rate upon the test set is a maximum. There is no guarantee that the best network performance with respect to the network's sum squared error (upon either the training set or test set data) reflects the set of weights yielding the best classification performance. Indeed, this is a limitation of the sum squared error cost function used by the backpropagation algorithm. An error measure embodying some measure of classification performance would capture the requirements of a pattern recognition neural network more appropriately.

Hidden Layer Nodes. The optimum number of hidden layer nodes is difficult to establish, and is strongly dependent on the nature of the data. It is not likely that this issue will be resolved for the general case, since each problem demands different capabilities of the network. Having said this, choosing the proper

network size is important. If the network is too small, it will not be capable of forming a good model of the system. Conversely, if the network is too large, then it may be *too capable*. That is, the network may be able to implement numerous solutions that are consistent with the training data, but most of which are likely to be poor approximations to the actual problem. In pattern recognition, this means that the training set may be easily discriminated, but the boundaries constructed may generalize poorly to the test set. The optimum size is that which would enable the network to capture *only* the underlying structure of the data.

With little or no knowledge of the underlying intricacies of the data, however, one must determine the network size by trial and error. A methodical approach is recommended. One may start with a very small network and gradually increase the size until performance tends to level off, training each network independently. Other approaches include *growing* [Tukey74] and *pruning* [leCun90] algorithms, which add or delete nodes as needed, respectively. Insofar as determining an upper bound on the number of hidden layer nodes, this number should be much less than the number of training samples, or the network will simply “memorize” the training samples, resulting in poor generalization.

Numerous methods have been proposed to improve network generalization, including network pruning [le Cun90], weight sharing [Waibel89], and weight decay [Hanson89] techniques. In addition to network architecture, generalization is affected by the number of patterns (and how well they represent the problem), and the complexity of the problem at hand.

2.3.2.4 The Capabilities of Artificial Neural Networks as Classifiers

Recall the earlier discussion on statistical pattern recognition. The objective is to approach the optimal performance of a Bayes classifier, in which the *a posteriori* probabilities $P(y_k | \mathbf{x})$ are known: given an input pattern \mathbf{x} , what is the probability that the output is class y_k ? Selection of the class with the highest probability is guaranteed to provide classification with the lowest probability of error. Although Bayes' decision rule is quite simple, it is often difficult to implement because the *a posteriori* probabilities are usually unknown. This means that they must be *estimated*. There are numerous ways of estimating these probabilities. For example, one may use the *a priori* probabilities and density functions by using Bayes' rule:

$$P(y_k | \mathbf{x}) = \frac{P(y_k)p(\mathbf{x} | y_k)}{p(\mathbf{x})}. \quad (2.19)$$

This approach characterizes most statistical pattern recognition methods [Tou74].

It is possible, however, to estimate the *a posteriori* probabilities directly; this is the methodology of a pattern recognition neural network. The problem is actually one of function approximation, which is what most supervised learning algorithms set out to do. Given a set of training examples, we wish to train an estimator $\theta_k(\mathbf{x}, \mathbf{w})$ to approximate $P(y_k | \mathbf{x})$ for each class $k = 1, \dots, K$. The vector \mathbf{w} represents the parameters of the estimator to be determined by the training process: the weights of the MLP ANN. In an ideal situation, the training set would consist of the input-output pairs $(\mathbf{x}, P(y_k | \mathbf{x}))$. The *a posteriori* probabilities are seldom available, however, and we have no choice but to use the class labels y_k .

Fortunately, it can be shown that, when a mean squared error criterion is used in training, and 0s and 1s are used as the target outputs (in place of the $P(y_k | \mathbf{x})$), the optimal solution for the parameters of $\theta_k(\mathbf{x}, \mathbf{w})$ are exactly the same as if the true *a posteriori* probabilities were used as the target outputs² [Duda73]. Thus, when the backpropagation algorithm is used, the MLP ANN can learn the best mean-squared-error approximation to the *a posteriori* probabilities. The sigmoid nonlinearities automatically bound the output to between 0 and 1, and it has been shown that the MLP outputs for any given pattern tend to sum to one (indicative of a probability density function) without any special constraints [Denker91]. Thus, not only does a properly trained MLP ANN offer the classification ability of Bayes' classifier, it may be capable of providing a measure of the confidence in the decision as well. In order to fulfill this very desirable capability, the network must be the right size for the problem, and the number of training examples must be sufficiently large to provide good generalization.

2.3.3 Other Classifiers

Multilayer perceptron neural networks and linear discriminant analysis are simple, yet effective pattern classifiers and for this reason, they are widely used. Many alternative approaches to pattern recognition exist however, and ultimately, the best classifier depends on the nature of the data to be classified. What follows is a brief overview of other important classifiers, which offer a slightly different approach to the classification problem.

² Similar results exist for other criterion functions as well [Duda73].

2.3.3.1 Distance Classifiers

Pattern classification by distance functions is perhaps the simplest and most intuitive approach to the problem. The motivation for using distance functions as a classification tool follows naturally from the obvious notion that the *similarity* of pattern vectors may be measured by their *proximity*. Pattern classification by distance functions can be expected to yield satisfactory results only when the classes tend to be well-clustered. This is illustrated in Figure 2.7, which depicts two datasets, one of which is easily classifiable by the proximity concept, and one that is not.

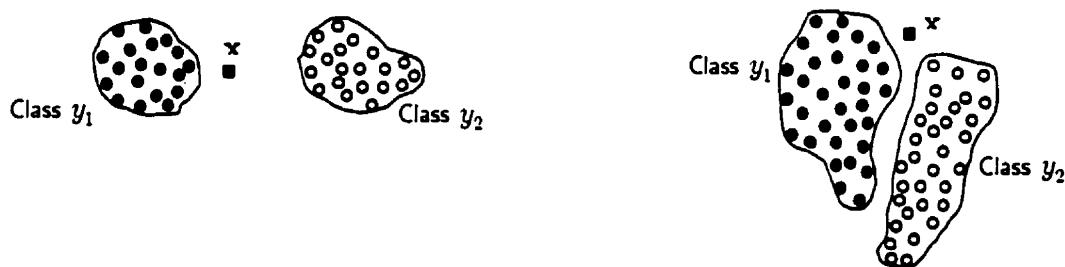


Figure 2.7 – Two datasets which demonstrate the applicability of distance classifiers. On the left, the novel pattern x is easily assigned to a class; on the right, the choice is not as clear despite the fact that the classes are linearly separable.

Since the proximity of an unknown pattern to the patterns of a known class serves as a measure for its classification, these approaches are termed *minimum-distance classifiers*. Sometimes, the patterns of each class tend to cluster tightly about a typical or representative pattern for that class, such that the pattern variability is well behaved. In this case, the simplest minimum-distance classifier can be effective; this is one based upon a *single class prototype*. Consider K pattern classes and assume that one may represent these classes by a set of prototype patterns $\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_K$. The Euclidean distance between a given pattern vector \mathbf{x} and the i^{th} prototype vector is

$$D_i = \|\mathbf{x} - \mathbf{z}_i\| = \sqrt{(\mathbf{x} - \mathbf{z}_i)^T (\mathbf{x} - \mathbf{z}_i)}. \quad (2.20)$$

A minimum-distance classifier computes the distance from an unknown pattern \mathbf{x} to the prototype of each class, and assigns the pattern to the closest class. This single prototype is often the mean vector of the pattern vectors within each class.

Suppose that instead of being representable by a single prototype pattern, each class is better characterized by *several prototypes*. That is, each pattern of class y_i tends to cluster about one of the prototypes $\mathbf{z}_i^1, \mathbf{z}_i^2, \dots, \mathbf{z}_i^{N_i}$, where N_i is the number of prototypes representing class i . Under these conditions, we can construct a multiple prototype minimum-distance classifier, where the distance function between a pattern vector \mathbf{x} and the i^{th} class is:

$$D_i = \min_l \|\mathbf{x} - \mathbf{z}_i^l\|, \quad l = 1, 2, \dots, N_i. \quad (2.21)$$

That is, the distance is the smallest of the distance between \mathbf{x} and each of the prototypes of class i .

It is easily shown [Tou74] that the decision boundaries for minimum-distance classifiers are the perpendicular bisectors of the lines joining the prototypes of different classes. Therefore, minimum-distance classifiers are a special case of linear classifiers, in which the decision boundaries are constrained to have this property.

Although the ideas of prototypes and Euclidean distances are geometrically attractive, they are not limiting concepts in the definition of minimum-distance techniques. Consider a set of sample patterns of known class membership $\{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_P\}$, where it is assumed that each pattern belongs to one of the classes

y_1, y_2, \dots, y_K . We may define a *nearest-neighbor* (NN) classification rule which assigns a pattern \mathbf{x} to the class of its nearest neighbor, where we say that $\mathbf{s}_i \in \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_P\}$ is a nearest neighbor to \mathbf{x} if

$$D(\mathbf{s}_i, \mathbf{x}) = \min_p \{D(\mathbf{s}_p, \mathbf{x})\}, \quad p = 1, 2, \dots, P \quad (2.22)$$

where D is any distance measure defined on the pattern space.

We may call this scheme the 1-NN rule since it employs only the class membership of the nearest neighbor to \mathbf{x} . We can, in this spirit, define a k -NN rule which consists of determining the k nearest neighbors, and classifying \mathbf{x} according to the most prevalent class in this group.

One of the drawbacks of the k -NN method is that, in order to provide a sufficiently rich set of exemplars, it is necessary to store a large set of sample patterns of known classification. In addition, the distances from each pattern to be classified to all the stored samples must be computed for classification. This represents a severe computational burden for large datasets. In the interest of reducing the storage and computational requirements, algorithms have been devised to reduce the size of the labeled set [Hart68][Fukunaga90].

2.3.3.2 Classification and Regression Trees

Tree-based classifiers can be viewed as a nonlinear discriminant function approach. The tree-based algorithm proceeds by constructing a series of simple greedy splits of the data into subgroups. At each node, the split which best classifies the signals in the left and right branches is selected. Each subgroup is then split recursively, so that the resulting classifier has an hierarchical binary tree structure. Splitting is

continued until nodes become “pure”, *i.e.*, they contain only one class of signals, or become “sparse”, *i.e.*, they contain only a few signals. Class assignments are made at the terminal nodes, usually by a majority vote of the samples belonging to that node.

After growing the initial tree, a pruning process is usually applied to eliminate unimportant branches, which helps combat “overtraining.” The details of splitting, growing and pruning rules can be found in [Breiman84]. This approach is particularly useful in problems where the input patterns contain a mixture of symbolic and numerical data. It also provides a rule-based interpretation of the classification protocol (the decisions made at each node determine the rules). Although classification and regression trees do not assume any parametric model for the data distributions, they are still vulnerable to the curse of dimensionality, and pose a significant computational burden for high-dimensional input spaces.

2.3.4 Summary: Pattern Classification

This section has offered a brief synopsis of important pattern classifiers in use today. This overview is by no means comprehensive in its scope, but rather, is intended as a reasonably diverse representation of classifiers that may be considered for a given application.

Generally speaking, most classifiers can be designed to yield near-optimum performance for most problems. The main differences between them involve the number of free parameters (which govern generalization and training set size issues) and the geometry of the decision boundaries which they construct. Other

differences lie in areas such as their time complexity of learning, their computational and storage requirements, and their robustness.

The primary focus of this work however, is upon the efficacy of signal representation for classification, not upon the classifiers themselves. Of greater interest here is how well the measured data can be preprocessed, so as to simplify the classification task. For this reason the desirable properties of a classifier here are that it be simple, well understood, easily interpreted and easily implemented. Correspondingly, two classifiers will be used throughout: a linear discriminant analysis, and the multilayer perceptron.

The advantages of the LDA are that it is conceptually very simple, and training is done in one pass: a one-time solution of an eigensystem specifies the discriminant boundaries. For datasets that are approximately normally-distributed, the LDA performs very well. The limitations of the LDA include, obviously, the fact that the discriminant bounds are linear, and by the constraints imposed by the assumption of Gaussianity. The LDA can be distracted by noisy data, and is therefore sensitive to the effects of the curse of dimensionality. MLP networks trained with the backpropagation algorithm offer good functionality with low complexity, insofar as neural networks go. Training may be slow for some datasets, but the MLP offers the ability of constructing arbitrarily complex decision boundaries. To some degree, the MLP can suppress noisy or irrelevant data, if properly trained. In this situation, its generalization ability surpasses that of the LDA.

The focus in this chapter now turns to the challenges of feature extraction.

2.4 Feature Extraction

2.4.1 The Importance of Feature Extraction

The preceding section explained the various techniques available for pattern classification. Before a pattern classifier can be properly designed or effectively used however, it is necessary to consider the feature extraction and data reduction problems. Although feature extraction should be considered before a classifier is designed, a greater appreciation of the importance of feature extraction is gained when the order of presentation of these two topics has been reversed, as has been done here.

Based upon empirical observation, multivariate data occupying N dimensions (\Re^N) are almost never N -dimensional [Scott92]. That is, the *underlying structure* of data in \Re^N is almost always of dimension lower than N . Therefore, the input space may be usefully partitioned into subspaces of signal and noise. Of course, this partition is rarely precise, but the goal is to eliminate a significant number of dimensions to obtain an efficient representation of the underlying structure. In the context of pattern classification, feature extraction consists of choosing those features which are most effective for preserving class separability.

It is fair to say that feature extraction plays a central role in pattern recognition. In fact, the selection of a feature set which accommodates the difficulties in the selection or extraction process, and at the same time results in acceptable performance, is the most challenging task of classification.

Feature extraction methods can be divided into two groups: *statistical* (or *decision-theoretic*) and *structural* (*syntactic*). *Statistical feature extraction* has, by far, received the most attention. This is because statistical methods lend themselves to direct mathematical description and machine implementation. Amongst the significant contributions to statistical feature extraction are the orthogonal transform methods (the FFT, WT, singular value decomposition, etc.).

Syntactic or structural methods, on the other hand, explicitly utilize the structure of patterns. A signal is described in hierarchical fashion as a composition of “primitives” or “elementary building blocks” (e.g. peaks/valleys or energy subbands). Specifically, this approach associates the description of signals with a formal *language theory*. The language analogy is fundamental:

| Syntactic Signal Description | Language |
|-----------------------------------|---------------------------------|
| “Primitives” of signals | Words |
| Signals | Sentences |
| Structural description of signals | Syntax or grammar of a language |

Statistical approaches assume very little about the structure of the data under consideration, while syntactic methods attempt to explain it as best they can. Of course, the existence of a recognizable “structure” is essential for the success of the syntactic approach. It is natural for humans to utilize structural feature extraction as the basis for classification and recognition, but it has proven extremely difficult to allow machines to reproduce this type of processing. Consequently, syntactic pattern recognition applications have been confined to a specific subset of problems such as image processing, which are characterized by recognizable shapes and edges that may be transformed into a more convenient

mathematical representation. It is beyond the scope of this work to investigate whether the MES may be modeled as an hierarchy of primitives and grammar.

It is important to note, however, that transform-based methods of feature extraction do embody some the precepts of structural methods. The parametric nature of a transform necessarily implies a set of “primitives” (the transform basis functions) which constitute the “atoms” of the signal description. The transform itself is the “grammar” which binds the basis functions *via* analysis and synthesis. The ability of a transform method to capture a pattern’s structure depends entirely upon how appropriately the basis functions model the pattern.

2.4.2 Feature Extraction for Classification

When we have two or more classes of data, the goal of feature extraction is to choose those features which are most effective for preserving class separability. This section will explore the issues in selecting a feature set that accurately represents the physical process of interest (here, the transient myoelectric signal), and one that is ready for practical implementation.

Approach 1: *Construct an “ideal” feature set using the probability distribution of the training data.*

Recall the definition of the classification problem given in Section 2.1: the input signal space is $\mathcal{X} \subseteq \mathbb{R}^N$ and the response space is $\mathcal{Y} = \{y_1, \dots, y_K\}$, and (\mathbf{x}, y)

represents the input pattern / class label pair. Class separability depends not only on the class distributions $p(\mathbf{x} | y_k)$, but also on the classifier to be used; the optimum feature set for a particular classifier may not be optimum for another. Assume for now that we seek the optimum feature set with respect to the Bayes classifier; this will result in the minimum error for the given distributions. From Section 2.3.1 we have that, for a K -class problem, the Bayes classifier consists of selecting the largest of the *a posteriori* probabilities $P(y_k | \mathbf{x})$, for $k = 1, \dots, K$. These may be determined directly, or by using the conditional probabilities $p(\mathbf{x} | y_k)$ in conjunction with Bayes' theorem.

We may therefore identify the *ideal* feature set as that which comprises the K *a posteriori* probabilities $P(y_k | \mathbf{x})$. Actually, since these probabilities sum to unity, only $K - 1$ of the *a posteriori* probabilities are needed to specify the Bayes classifier. In this case, class separability is *equivalent* to the probability of error of a Bayes classifier, which is the best we can expect. Therefore, theoretically speaking, the Bayes error is the optimum measure of feature effectiveness.

It is seldom possible to construct a true Bayes classifier, however. The probabilities $P(y_k | \mathbf{x})$ are not known in practice, and they are often extremely difficult to estimate from available training samples without severe bias or variance, especially if the dimension of the input space, N , is high. The Bayes error is just too complex and useless as an analytical tool to extract features systematically.

Whereas the *a posteriori* probabilities $P(y_k | \mathbf{x})$ constitute the *ideal* feature set, one must seek a feature set that is *best* or *optimum* with respect to some simpler

criterion which may yield systematic feature extraction algorithms. A generalized feature extraction method may be represented as a map $f: \mathcal{X} \rightarrow \mathcal{F}$, such that $\mathcal{F} \in \Re^M$ is the M -dimensional *feature space*, where $M \ll N$. Its goal is to reduce the dimension of the classification problem by extracting a feature set which retains important information and discards that which is irrelevant for the purpose of classification. It is much easier to build classifiers and to analyze the pattern data (numerically and graphically) in a lower dimension.

The key question is: *how is the “best” feature set (or feature extractor, f) determined?* The qualifier “best” is largely application-dependent. A feature set that is optimal for signal compression may be less than adequate for pattern classification. Consider the (unlikely) situation where the physical and mathematical properties of a process are precisely known. In this case, the *intrinsic dimension* of the data is known, and a feature set may be determined analytically that best describes the process (in a dimension equal to the intrinsic dimension). Here, we have the desirable properties of (a) preservation of information, and (b) efficiency of description.

It is rare, however, that a mathematical model exists for a physical signal that would allow derivation of a feature set in this way. It is customary, therefore, to adopt exploratory techniques that have limited assumptions about the structure of the data. In this manner, a feature set is determined by procedures set in the context of a criterion that evaluates the feature set’s efficacy. This leads to the next strategy for signal classification:

Approach 2: Determine the feature set that best discriminates a given set of signals, according to some criterion function.

A feature set may be considered optimum in some sense if it maximizes (or minimizes) a chosen criterion function. There are two fundamentally different approaches to determining the best feature set for a given application; these will be termed *feature selection* and *feature projection* in this work

Feature selection methods attempt to reduce the number of variables by selecting the best *subset* of the input space for class discrimination, according to some criterion. Usually, class membership is used in the determination of discriminant power, and these methods may be considered *supervised*. A desirable trait is that the original identity of the features is maintained, which is useful in interpreting the utility of individual features. The main disadvantage of feature selection methods is that some of the original features, and thus, the information that they convey, must be discarded completely. This is a problem if class separability information is broadly dispersed throughout the original feature set.

Feature projection methods aim to describe the data as concisely as possible, with a minimal loss of information. Essentially, the optimal combination (or projection) of the original feature space into a smaller set of new features is sought, with respect to a chosen criterion. Principal components analysis (PCA) [Haykin94] and projection pursuit [Huber85] are examples of feature projection methods. The resulting feature set is *ranked*, according to the importance assigned by the criterion function, and the best $M < N$ of the features are chosen as a representation. For example, having performed a PCA, the resulting principal components are ranked according to the information they contribute to the data,

in a mean square error sense. A subset of the largest principal components then may be chosen to define the reduced feature set.

Feature projection methods are usually *unsupervised*; they operate on the aggregate of the data, and do not include class membership in their criteria. The advantage offered by projection-based methods is that they produce the best linear projection, using *all* of the original features.

The optimizing characteristics of selection and projection-based methods are alluring: it would be intuitively satisfying to say that a feature set is the result of the best subset or projection of the measured data. These methods have their limitations, however. Projection-based methods may become confused by data clouds that have many isotropically distributed clusters [Freidman74]. Similarly, it has been shown that feature selection methods do not work well on high-dimensional data, especially those which have excessive redundancy and dispersion of information [Tate96].

A direct implication of these observations is that these methods should not be applied directly to the measured signals. It is preferable to perform some form of preliminary feature extraction before applying either projection or selection-based methods. Therefore, these methods must be considered an *adjunct* to the process of feature extraction: an additional stage of *dimensionality reduction* following initial feature extraction. We may loosely refer to the combined actions of feature extraction and dimensionality reduction as the *signal representation stage*. To illustrate the role of dimensionality reduction, a refined version of the classification process is shown in Figure 2.8.

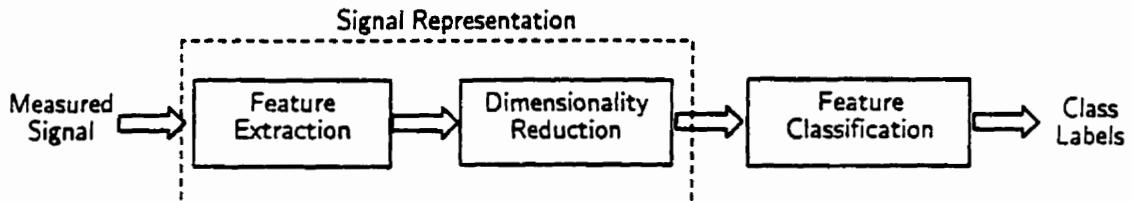


Figure 2.8 – A refined model of the classification problem, including the stage of dimensionality reduction.

Although there is inevitably some overlap in function, the roles of feature extraction and dimensionality reduction may be generalized as:

Feature Extraction:

Transform the data so as to emphasize the invariant structures in the data, while rejecting irrelevant data (“noise”). Feature extraction provides a potentially more revealing characterization of the data; for classification applications, it must provide good separation between classes.

Dimensionality Reduction:

Determine the best *subset* or *combination* of features for the purpose of classification. Reducing the dimensionality of the problem simplifies the task of the classifier, and alleviates the generalization problems due to the curse of dimensionality. Dimensionality reduction for signal classification will be treated in detail in Section 2.6.

Let us return again to the problem of feature extraction. It has been explained that, in general, it is impossible to construct an “ideal” feature set consisting of Bayesian *a posteriori* probabilities. If we are to determine the “best” feature set subject to some criterion function, some systematic method of feature selection or feature projection must be applied. These methods perform poorly, however, when

acting directly on the measured signals, especially those that are high-dimensional. It is essential then, that an effective form of feature extraction precede the dimensionality reduction stage. For signals of a transient nature, such as the transient MES patterns here, the feature extraction stage must capture the temporal structure present in the raw data. Time-frequency analysis provides a powerful and structured framework for transient signal representation, localizing information simultaneously in both domains. This is especially important for transient signals of short duration, where the information in time and frequency may be closely intertwined. Fast transforms exist for most time-frequency methods, and most are readily applicable to specialized digital hardware implementation. The third approach, therefore, may be stated as:

Approach 3: Determine the time-frequency representation that best represents a given set of signals for the purpose of classification.

The efficacy of time-frequency based feature sets for transient MES classification is the major course of this investigation.

2.5 Time-Frequency Representations

The primary goal of signal analysis is to extract information from a signal, relevant to a particular application. Time-frequency representations (TFRs) combine time-domain and frequency-domain analyses to yield a potentially more revealing picture of the temporal localization of a signal's spectral characteristics. TFRs have found utility in virtually every area of signal processing, including signal compression, coding, filtering, noise suppression, detection, classification, and visualization. TFRs may be divided into two groups³ by the nature of their transforms: *linear methods* (including the short-time Fourier transform and the wavelet transform) and *quadratic methods* (of which the Wigner-Ville distribution is fundamental).

The concept central to *linear methods* is that of decomposing a signal into elemental pieces or *time-frequency atoms*. That is, a signal $x(t)$ can be represented as an N -member synthesis:

$$x(t) = \sum_{k=0}^{N-1} a_k \phi_k(t) \quad (2.23)$$

where $\phi_k(t)$ are the time-frequency atoms (or basis functions), and a_k are the corresponding coefficients. Desirable properties of the basis functions include computational efficiency, orthogonality, and good time-frequency localization.

³ Parametric methods, such as autoregressive (AR) and autoregressive moving average (ARMA) models, have been adapted for time-frequency analysis. These methods are quite different than linear and quadratic transform methods, however, and will not be directly addressed here.

The time-frequency localization of these basis functions and the amplitude of their coefficients describe the signal's TFR.

Quadratic methods are based upon estimating an instantaneous power spectrum (or energy distribution) using a bilinear operation on the signal $x(t)$ itself. Although quadratic methods are too computationally intense for real-time application, their high resolution provides valuable insight as a visualization tool.

The discussion concludes with a specification of the desirable properties that a TFR must possess in the context of feature extraction for signal classification.

2.5.1 The Short-Time Fourier Transform

Most transforms, in their original form, assume that the signal under consideration is *stationary*; that is, the statistical properties do not evolve in time (for a complete treatment of the notion of stationarity, see [Papoulis85]). The most widely-used and studied of these is the Fourier transform [Fourier88]:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-2j\pi ft} dt. \quad (2.24)$$

The analysis coefficients $X(f)$ denote the distribution of the signal in the frequency domain for the entire record; that is, with no temporal resolution. This is appropriate if the signal is comprised of a few stationary components (e.g. sinusoids), but any abrupt changes in time in a nonstationary signal $x(t)$ would spread out over the entire frequency axis of $X(f)$. The basic Fourier transform is obviously not suited for transient signals.

The usual approach to introduce some time dependence to the Fourier transform is to introduce the concept of “local frequency”; the Fourier transform “looks” at the signal through a window that is localized in time. The *short-time Fourier transform* (STFT) was first adapted by Gabor [Gabor46] to define a two-dimensional time-frequency representation. Consider a signal $x(\tau)$ and assume it is stationary when seen through a window $g(\tau - t)$ of limited extent, centred at the location t . The Fourier transform of the windowed signal $x(\tau)g^*(\tau - t)$ yields the STFT:

$$\text{STFT}(t, f) = \int x(\tau)g^*(\tau - t) e^{-2\pi ft} d\tau \quad (2.25)$$

which maps the signal into a two-dimensional time-frequency plane (t, f) .

It is important to consider the bounds on *temporal resolution* Δt and *frequency resolution* Δf of the STFT. The time-frequency resolution depends entirely upon the choice of the window $g(t)$. Given a window function $g(t)$ and its Fourier transform $G(f)$, a commonly used measure of frequency resolution is the *mean-square bandwidth* [Marple87]:

$$\Delta f^2 = \frac{\int f^2 |G(f)|^2 df}{\int |G(f)|^2 df}. \quad (2.26)$$

The corresponding measure of temporal resolution is the *mean-square time width*:

$$\Delta t^2 = \frac{\int t^2 |g(t)|^2 dt}{\int |g(t)|^2 dt}. \quad (2.27)$$

The resolution in time and in frequency cannot be arbitrarily small; their product is lower bounded by the *time-bandwidth uncertainty principle* or the *Heisenberg inequality* [Papoulis85]:

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi}. \quad (2.28)$$

This means that one must trade time resolution for frequency resolution, or *vice versa*. A Gaussian window is the only window that meets the Heisenberg bound with equality. The STFT using this particular choice of analysis window has been termed the *Gabor transform* [Gabor46]. In general, a window is chosen to meet the needs of a particular application, which may include time-frequency localization, computational efficiency, and suppression of spectral leakage.

A family of time-frequency atoms with uniformly bounded Heisenberg product may be represented by information cells of equal area. A *basis* of such atoms corresponds to a *cover* of the time-frequency plane by rectangles; an *orthonormal basis* may be depicted as a cover by *disjoint* rectangles. The choice of basis determines the aspect ratio of the rectangular time-frequency cells. For example, the *standard basis* (or *Dirac basis*) has optimal time localization but no frequency localization, while the Fourier basis has optimal frequency localization but no time localization. These are depicted in Figure 2.9.

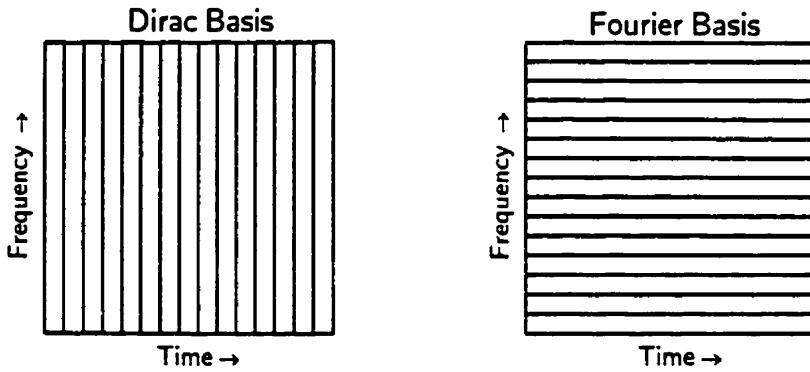


Figure 2.9 – Dirac and Fourier bases provide a disjoint tiling of the time-frequency plane.

For a given physical signal, the most appropriate basis is a compromise between time and frequency resolution. This tradeoff is illustrated in Figure 2.10, which depicts the continuous spectrogram⁴ of a signal composed of two sinusoids of different frequency, and two Kronecker delta pulses located at different locations. A Gaussian window was used, providing the best dual localization in time and frequency. On the left, the spectrogram was computed using a relatively narrow window, resulting in good temporal localization, but poor frequency resolution. On the right, a Gaussian window twice as wide yields much better frequency localization, at the expense of temporal resolution.

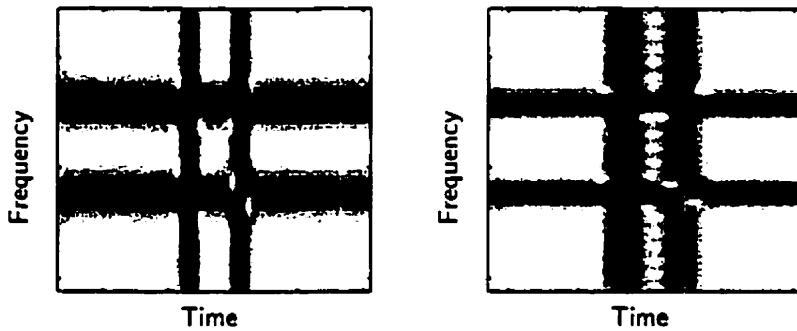


Figure 2.10 – The continuous spectrogram of a signal composed of two sinusoids and two Kronecker delta functions. On the left is a spectrogram using a narrow analysis window; on the right, the spectrogram of the same signal, using a broader analysis window.

⁴ The spectrogram is usually preferable to the STFT for visualization, since the STFT is, in general, complex valued. The spectrogram is the squared magnitude of the STFT, approximating an energy density distribution in the time-frequency plane, while discarding phase information. The spectrogram is therefore actually a quadratic time-frequency representation. This means that *cross terms* will exist between spectral components that are not present in the STFT. These interference terms are evident between the vertical bands representing the Kronecker delta pulses in Figure 2.10. The nature of cross terms in quadratic representations is discussed in Appendix B.

The continuous STFT, however, is a highly redundant representation. For practical application (terms of memory requirements and computational complexity), the STFT must be discrete in time and in frequency. Consider a signal sampled at a frequency f_s (its sampling period is T_s). If the window $g(t) \equiv g(iT_s)$ is L samples in duration, then the discrete Fourier transform (DFT) at time $t = iT_s$ is:

$$X(mF) \equiv X[m] = \sum_{i=1}^{L-1} x[i] e^{-2\pi i (mF)(iT_s)}, \quad (2.29)$$

where $F = \frac{1}{LT_s}$ is the *frequency sampling step size*⁵. The discrete STFT consists of a series of DFTs, indexed with respect to T_s and F :

$$\text{STFT}[k, m] \equiv \text{STFT}(kT_s, mF) = \sum_{i=1}^{L-1} x[i] g[i-k] e^{-2\pi m i / L}. \quad (2.30)$$

In general, the time samples are spaced K samples apart. In continuous time, this *temporal sampling step size* is $T = K \cdot T_s$. Thus, the STFT may be indexed by its time-frequency sampling steps T and F :

$$\text{STFT}[n, m] = \text{STFT}(nT, mF) = \sum_{i=1}^{L-1} x(iT_s) g(iT_s - nT) e^{-2\pi m i / L} \quad (2.31)$$

If $K = 1$, then the STFT is computed at every sample in time. If $K = L$, then successive analysis windows are non-overlapping.

Figure 2.11 shows the segmentation in the time-frequency plane of the sampled STFT. As imposed by the temporal and frequency sampling steps, each cell has a

⁵ The notation for a continuous signal employs round brackets surrounding an absolute (time or frequency) argument. The equivalent discrete signal uses square brackets surrounding an index with respect to the time or frequency step. That is, $x(iT_s) \equiv x[i]$ and $X(mF) \equiv X[m]$.

temporal width of T and a frequency height of F . A single coefficient represents the magnitude of the TFR within the cell boundary in the time-frequency plane.

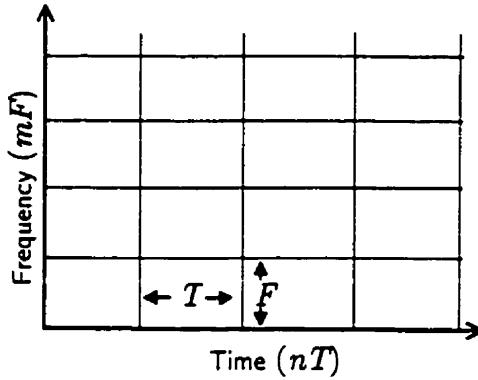


Figure 2.11 - The tiling in the time-frequency plane of the STFT.

If the window $g(t)$ is well localized in time and frequency, then the STFT may be expected to indicate the energy density of a signal at the time-frequency location (nT, mF) . Consider now a cell in the time-frequency grid: the area is $TF = KT \cdot \frac{1}{L}T_s = K/L$. If $TF = 1$ (or equivalently, if $K = L$), the TFR is said to be *critically sampled*, and the number of TFR coefficients equals the number of samples in the signal (assuming a bandlimited signal sampled at the Nyquist rate). Moreover, this is the largest grid that will allow the signal to be perfectly reconstructed. If $TF < 1$, the grid is said to be *oversampled*, and the TFR is redundant (as is the case with the continuous STFT). This may actually be a desirable property in applications such as data visualization. If $TF > 1$, the TFR is *undersampled*; the grid is too coarse for the signal to be reconstructed.

For classification problems, one attempts to choose a window function $g[i]$, a window length L , and sampling step sizes (T and F) to provide a TFR that is sufficiently resolved to capture the time-frequency structure, but not *too* resolved,

which would yield an unwieldy number of coefficients. Consider a simple example: a linear chirp signal ($N = 256$) shown in Figure 2.12.

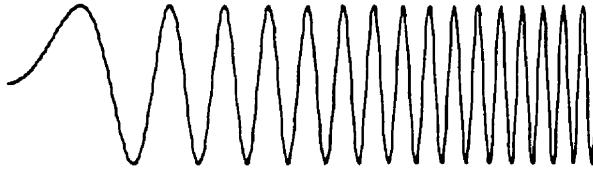


Figure 2.12 – A 256-point linear chirp signal.

Figure 2.13 shows a STFT analysis using a discrete Gaussian window. Three different combinations of window length L and sampling grid (T, F) were used for comparison. In Figure 2.13 (a), a window length of $L=32$ appears to capture the temporal transition in the chirp signal, but the time step of $K=1$ yields a TFR that contains too many cells to be useful as a classification feature set. Figure 2.13 (b) depicts the TFR using the same $L=32$ window, but with a nonoverlapping (critically sampled) grid. Here, the instantaneous frequency of the chirp is tracked reasonably well, and the energy is contained within a more manageable number (10-20) of TFR cells. In Figure 2.13 (c), a window length of $L=128$ and a time step of $K=128$ produce a TFR that does not have sufficient temporal resolution to track the instantaneous frequency of the chirp.

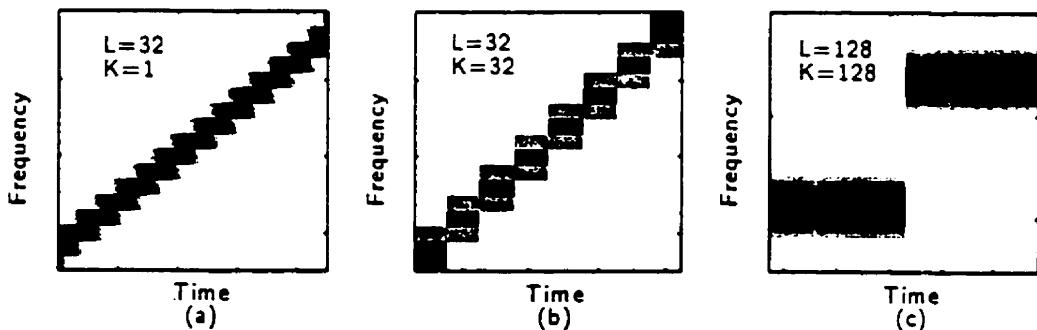


Figure 2.13 – Three alternative discrete STFT representations of a linear chirp signal. In (a), the time step ($K=1$) is too small, thereby generating many redundant TFR cells; in (b) the time step is equal to the window size, yielding a more parsimonious description. In (c), the window size is too large to track the instantaneous frequency of the chirp signal.

This is a very simple example, but it does illustrate the considerations that must be made when designing a TFR for the purposes of classifier feature selection. The

issues of resolution and tiling in the time-frequency plane will be revisited later, in the context of feature set selection.

To summarize, the STFT has many useful properties, including a well-developed theory, an easily understandable interpretation, and it may be computed very efficiently. The main drawback of the STFT is that even the most carefully chosen sampling grid is nonetheless constrained by the fact that each cell in the time-frequency plane must have an identical shape. Clearly, the energy distribution of physical signals is not, in general, conveniently localized in regions of fixed aspect ratio.

2.5.2 The Wavelet Transform

A fundamental property of the wavelet transform (WT) is that the time resolution Δt and the frequency resolution Δf vary in the time-frequency plane. Wavelet theory has developed as a unifying theory only recently through the work of Grossman and Morlet [Grossman84] and others, although similar ideas have existed since the turn of the century in pure and applied mathematics, physics, and electrical engineering. Daubechies [Daubechies88] and Mallat [Mallat89a] first linked wavelet theory to discrete signal processing.

The *continuous wavelet transform* (CWT) allows a variable coverage of the time-frequency plane. The transform is defined as:

$$\text{CWT}_x(\tau, a) = \frac{1}{\sqrt{a}} \int x(t) \psi^* \left(\frac{t - \tau}{a} \right) dt \quad (2.32)$$

where $\psi(t)$ is a prototype window referred to as the *mother wavelet*. The analysis determines the correlation of the signal with *shifted* (by τ) and *scaled* (by a) versions of the mother wavelet. The digital implementation of the CWT can be computed directly by convolving the signal with a scaled and dilated version of the mother wavelet. For a reasonably efficient implementation, an FFT may be applied to perform the convolution [Jones91].

In its discrete form, $a = a_0^j$ and $\tau = n \cdot a_0^{-j}$ where j and n are integers; this is referred to as the *discrete wavelet transform* (DWT). If we choose $a_0 \approx 1$ and $n \approx 0$, we are close to the continuous case. The choice of $a = 2^j$ and $\tau = n \cdot 2^{-j}$ is the most common choice; this is referred to as a *dyadic wavelet basis*. It allows much greater computational efficiency than the CWT; fast algorithms have been developed that are $O(n)$. The dyadic form also simplifies the constraints on the mother wavelet to achieve another desirable property: *orthogonality* of the analysis windows, which subtends an efficient representation. The dyadic DWT is almost always the chosen implementation because of its computational efficiency and the mutual orthogonality of its analysis windows (or subbands), and it will heretofore be referred to as simply the wavelet transform (WT) unless otherwise specified.

Because the mother wavelet may take many forms, the term *scale* is often preferred to *frequency* when using the WT. An analogy between scale and frequency is useful however, at this point, to illustrate the difference between the STFT and the WT. Consider a mother wavelet chosen to be

$$\psi(t) = g(t) e^{-2\pi f_0 t} \quad (2.33)$$

where $g(t)$ is the STFT analysis window, and f_0 is the fundamental frequency in the STFT analysis. This implies that the frequency analogous to the WT scale is

$f = af_0 = 2^j f_0$. As the scale is increased, however, the frequency resolution (or equivalently, the *bandwidth*) of the orthogonal WT analysis window is proportional to its centre frequency:

$$c = \frac{\Delta f}{f} = \frac{2^j}{2^j f_0}, \quad (2.34)$$

where c is a constant. Thus, the bandwidths of the analysis windows are spread *logarithmically* with respect to frequency, instead of *linearly* as with the STFT. This is illustrated in Figure 2.14.

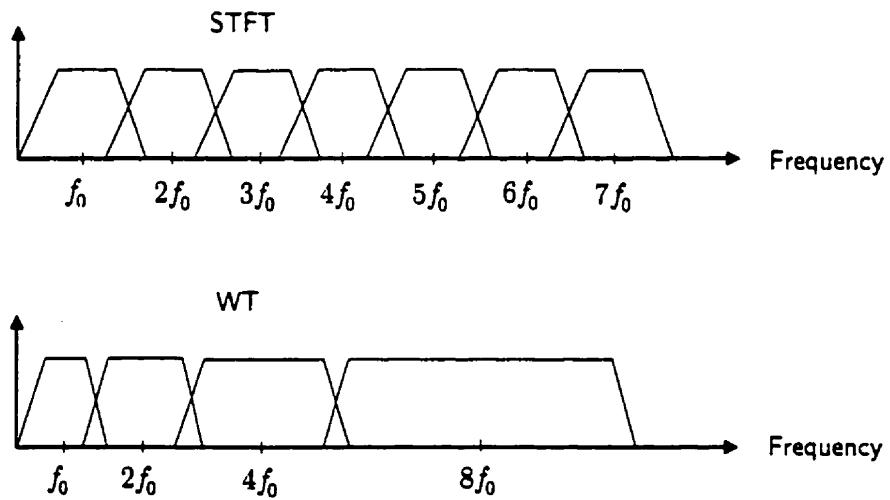
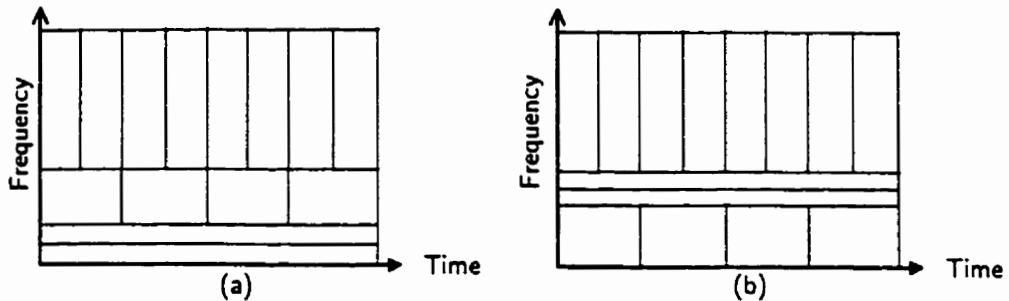


Figure 2.14 – Division of the frequency domain for the STFT and the WT.

The Heisenberg uncertainty principle still holds here, so the resulting tiling in the time-frequency plane is as shown in Figure 2.15(a).



**Figure 2.15 - The time-frequency plane tiling of
(a) a wavelet basis, and (b) an arbitrary wavelet packet basis.**

What ensues is that the time resolution Δt becomes arbitrarily good at high frequencies, and the frequency resolution Δf becomes arbitrarily good at low frequencies. For this reason, wavelet analysis is more effective than Fourier analysis when the signal of interest is dominated by transient behavior or discontinuities.

The *wavelet packet transform* (WPT) is a generalized version of the CWT and the DWT. The transform is redundant, allowing one of many orthogonal bases to be chosen [Coifman89][[Coifman90][Meyer93]. As a result, the tiling of the time-frequency plane is configurable: the partitioning of the frequency axis may take many forms to suit the needs of the application. This is illustrated in Figure 2.15(b).

Consider the same mixed signal used in the previous section, composed of sinusoids and Kronecker deltas. Figure 2.16(a) depicts the *scalogram* of the signal – the squared magnitude of the CWT. The tradeoff in time and frequency resolution is evident; the delta functions are much more resolved at higher frequencies, and the low-frequency sinusoid is more discernable than the high frequency sinusoid. The TFR of the wavelet transform is shown in Figure 2.16(b). It does a fair job of capturing the time-frequency characteristic of the signal, but the tiling prevents a

resolved representation of the high-frequency sinusoid and the low-frequency region of the Kronecker delta functions.

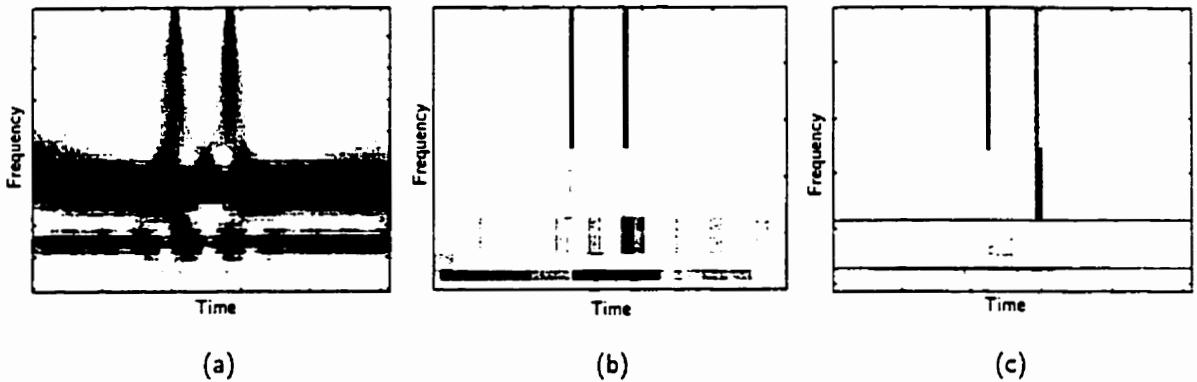


Figure 2.16 – The TFRs of wavelet-based representations: (a) the continuous wavelet transform (scalogram), (b) the discrete wavelet transform, and (c) the wavelet packet transform, based upon an entropy cost function.

Figure 2.16(c) demonstrates the ability of the wavelet packet basis to adaptively tile the time-frequency plane. The wavelet packet basis was specified using an entropy-based cost function to determine the *best* partition of the time-frequency plane [Wickerhauser94]. It is evident that the chosen wavelet packet basis localizes the multicomponent signal very well.

Wavelets and wavelet packets will be defined more rigorously in Chapter 3. Moreover, wavelet packet theory will be extended to yield a time-frequency tiling that is optimized for the problem of signal classification.

2.5.3 Quadratic Time-Frequency Representations

The main disadvantage of linear time-frequency transforms is that the time-frequency resolution is limited to the Heisenberg bound. This is due to the imposition of a local time window, $g(t)$: if this window is more resolved in time, the

frequency resolution suffers because the effective width of its Fourier transform $G(f)$ increases, and *vice-versa*.

An alternative approach to time-frequency representation is by means of a *quadratic* (or *bilinear*) *transform*. The fundamental form of all quadratic TFRs is the *Wigner Ville Distribution* (WVD) [Wigner71]. The WVD may be interpreted as a two-dimensional distribution of signal energy over the time-frequency plane. The absence of a windowing function yields a very high-dimensional TFR that offers superior resolution to any linear method. The details of quadratic TFRs are given in Appendix B.

Three factors hinder the application of the WVD as a method of feature extraction for pattern recognition:

1. **Computational Complexity.** Although some efforts have been made to reduce the computational load required to compute the WVD [Boashash87], the solutions do not nearly approach the computational efficiency of the FFT and fast wavelet transforms. Moreover, the algorithms usually include the overhead of computing the analytic signal, and performing two-dimensional filtering if cross-term smoothing is required. Further still, the task of performing systematic dimensionality reduction (such as PCA) on these high-dimensional TFRs may require enormous computational expense.
2. **Interference Terms.** The presence of cross-terms (see Appendix B) may obscure the relevant information for classification or “distract” the classifier by introducing irrelevant information. It may not be straightforward to discriminate amongst signal terms and cross terms with signals of physical origin because they are, in general, composed of many time-frequency

components. In this case, it is very likely that cross-terms and signal terms will overlap in the time-frequency plane.

3. Dispersion of Information. The primary objective of a feature set for classification is to reject irrelevant information and to represent important information in as few dimensions as possible. The superior resolution of the WVD may actually be to its detriment as a basis for a feature set, essentially “dispersing” information in a fine time-frequency grid. Important regions of the time-frequency plane may require the coefficients of many WVD cells to describe its boundary, whereas a less resolved – but appropriately tiled – representation may need only a few.

The WVD has found application in pattern recognition, but the instances have been few due to the challenging task of reducing the high dimension of the TFR down to a manageable size for classification. Marinovic [Marinovic85] has suggested that applying a singular value decomposition to the WVD can provide a robust feature set. Boashash [Boashash90] has described a method of classifying underwater acoustic signals by computing the cross-WVD. His method essentially performs a two-dimensional matched filtering operation between a signal’s WVD and a representative WVD template from each class. Both of these methods entail enormous computational load, and are unreasonable for real-time implementation using existing computing technology. For this reason, quadratic TFRs will be used in this work for visualization purposes only.

2.5.4 Time-Frequency Representations for Signal Classification

Recall the fundamental purpose of feature extraction for classification: this is to emphasize the important information in the measured signal, and to de-emphasize that which is irrelevant. This implies transforming the raw data into a domain that presents the information contained in the signal more clearly: a map which concentrates and localizes information. Time-frequency methods offer the ability of localizing the energy distribution of a signal in time and in frequency. The nature of the localization depends upon the method chosen, as described in the previous sections.

The utility of a TFR as a feature extractor for pattern classification lies in its ability to describe important structures in the time-frequency plane in a parsimonious manner. This requires an appropriate tiling of the time-frequency plane. Recall the nature of the time-frequency tiling for the transforms described in the previous sections.

- (i) The STFT segments the time-frequency plane into rectangles of fixed aspect ratio.
- (ii) The wavelet transform's TFR allows greater frequency resolution at lower frequencies and better time resolution at high frequencies.
- (iii) The wavelet packet transform permits an arbitrary segmentation of the frequency axis. The tiling is result of a basis selection procedure that optimizes a cost function chosen to evaluate the efficacy of the wavelet packet basis.

The time-frequency resolution of each of the above are bounded by the Heisenberg uncertainty. The WVD has unlimited resolution in time and in frequency, which may be compromised by applying a smoothing kernel to reduce

cross-term interference. As an example of TFR representation for classification, consider the (somewhat trivial) task of discriminating three localized sinusoids, each occupying a unique region in the time-frequency plane. Figure 2.17 depicts the TFRs subtended by a spectrogram (using a Hamming window of length 32, with non-overlapping segments), the wavelet transform, a wavelet packet transform (with a basis chosen to optimize an entropy cost function), and the WVD.

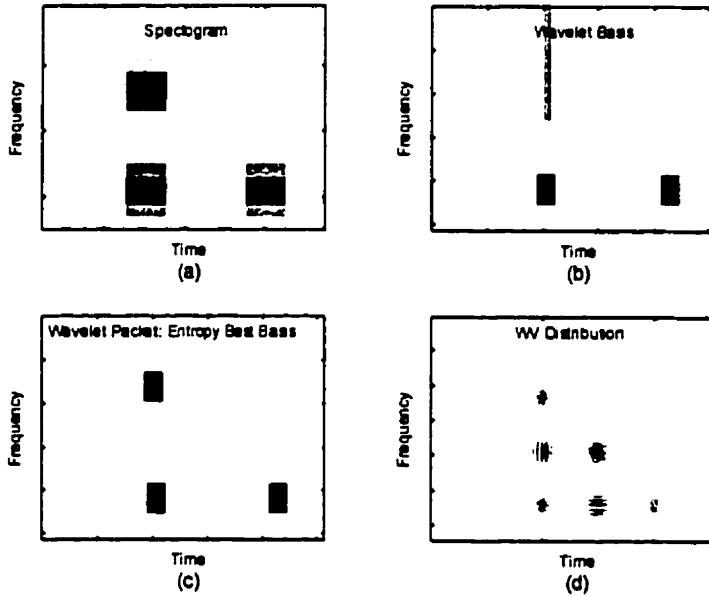


Figure 2.17 – The TFRs of three localized sinusoids; (a) the spectrogram, (b) the wavelet transform, (c) the wavelet packet transform, and (d) the WVD.

To be a good feature extractor, the TFR must cluster the information within each class, and to provide maximal discrimination of these clusters. When using the time-frequency plane as a feature space, it is imperative that the representation provide good localization (to prevent overlap of information that may provide discrimination) using as few TFR cells as possible (to simplify the role of the pattern classifier). The STFT does a fair job of localizing the three components, albeit with poor resolution. If the sinusoids were more proximal in time or

frequency, it is likely that some overlap would occur. The TFR of the WT demonstrates superior localization of the low-frequency sinusoids, but the frequency localization of the high-frequency sinusoid is somewhat ambiguous (at the expense of good time resolution). The wavelet packet transform, choosing the basis that minimizes the signal entropy, does a commendable job of localizing the three sinusoidal bursts, with most of the signal energy contained within only three TFR cells. The WVD localizes the information very well, but it requires many TFR cells to describe the time-frequency boundaries of each sinusoid, and the cross terms may confuse a pattern classifier.

Consider now the task addressed by this work: the use of the TFR as a feature set for classifying transient MES patterns. Figure 2.18 depicts the TFRs of two arbitrarily selected classes of one channel MES activity: elbow flexion and forearm supination. The left-hand column contains the data corresponding to a single elbow flexion; the right-hand column corresponds to forearm supination. Figures (a) and (f) depict the measured signals, sampled at 1000 Hz. Figures (b) and (g) represent the STFT processed with a non-overlapping hamming window of length 32. The same STFT analysis was repeated with a non-overlapping window of length 64, and these data are shown in Figures (c) and (h). Figures (d) and (i) depict the results of a wavelet packet analysis using an entropy-based basis selection cost function. For the sake of illustration, and to provide a more highly resolved picture of the time-frequency characteristics of the transient MES patterns, a Wigner-Ville distribution performed on the analytic signals is shown in Figures (e) and (j).

First, consider what information the highly resolved WVD TFR conveys. The elbow flexion seems to exhibit almost sinusoidal behavior and, correspondingly, the

WVD TFR indicates a rather narrow band of energy concentrated at a relatively low frequency. The amplitude builds gradually from the onset, peaks at about the centre, and decays only slightly toward the end of the record. The instantaneous frequency slowly increases, and disperses somewhat. The forearm supination waveform is more complex, but still fairly regular. It is noticeably more temporally compact than elbow flexion, and is characterized by the (almost) simultaneous existence of two dominant energy bands that experience a gradual decrease in instantaneous frequency.

It can be seen in Figures (b) and (g) that a STFT with a 32-point window does a fair job at localizing this behavior. The temporal resolution is sufficient to capture most time-dependent characteristics, but the frequency grid is too coarse to localize some of the band-limited information. Conversely, with a window of 64 points [Figures (c) and (h)] improved frequency discrimination comes at the expense of a coarse temporal grid. The wavelet packet transform's ability to adaptively tile the time-frequency plane is evident in Figures (d) and (i). Two narrow-band strips of 64 ms capture the essence of the elbow flexion in Figure (d), while a combination of cells with varying aspect ratio localize the information in the forearm supination signal in Figure (i).

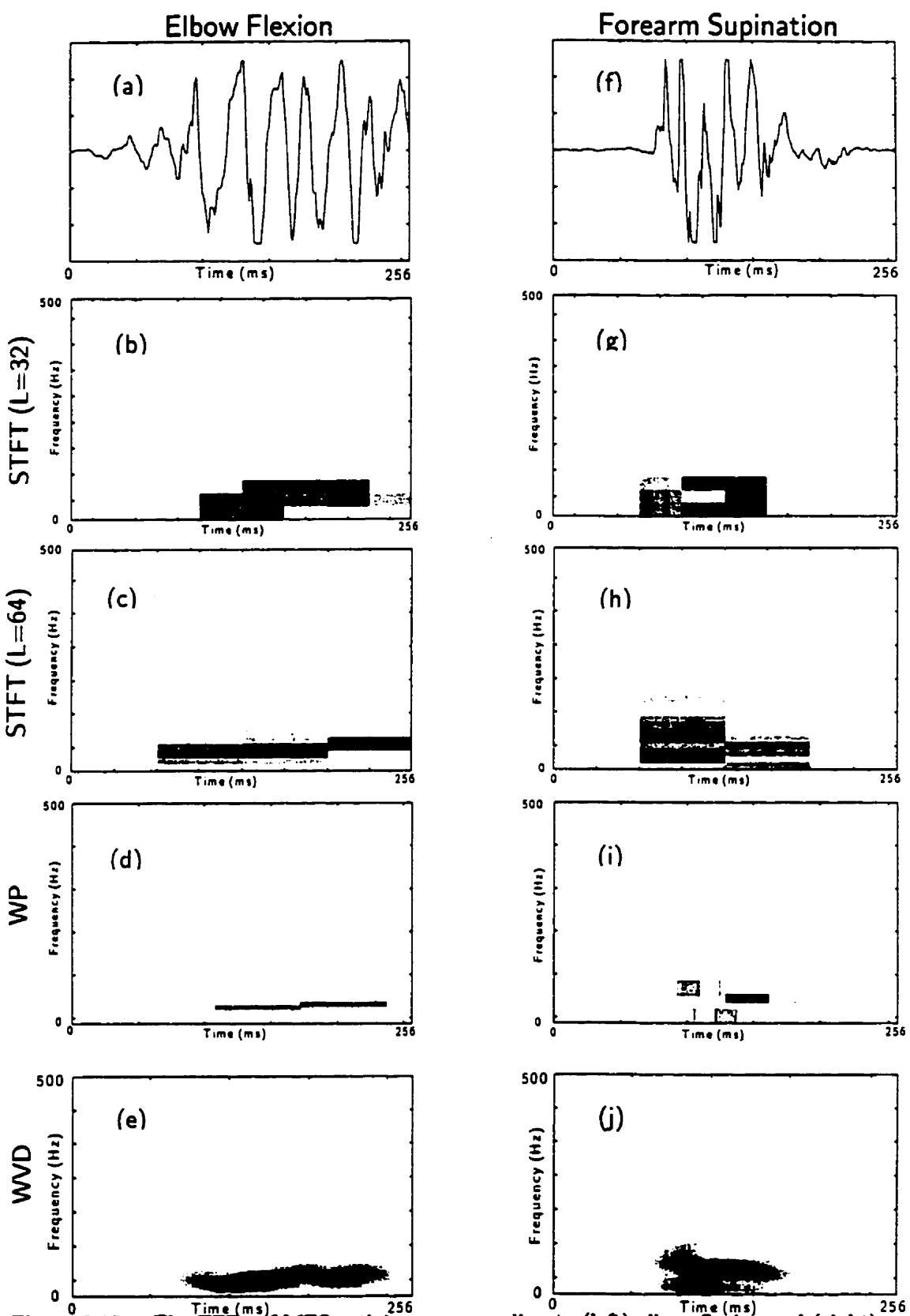


Figure 2.18 – The TFRs of MES activity corresponding to (left) elbow flexion and (right) forearm supination. Figures (a) and (f) – the measured signals; (b) and (g) – a STFT with window size=32; (c) and (h) – a STFT with window size=64; (d) and (i) – a WP transform; (e) and (j) – the WVD.

The question is: what is the relative capability of each method with respect to signal classification? The answer depends not only on how well a particular TFR method models a set of signals, but on how well it identifies inter-class distinctions, and rejects the intra-class variability. It is apparent that the wavelet packet transform has the ability to adaptively tile the time-frequency plane, thereby localizing salient energy clusters effectively. It should be noted that the wavelet packet basis selection (and therefore, the time-frequency tiling) done here acts upon each signal individually, using an entropy-based cost function. This method of basis selection is optimal for signal reconstruction (compression), providing the best localization of energy for a given signal, and also the best tiling for TFR visualization. This basis is appropriate for our brief inspection of the time-frequency properties of the MES here, in terms of providing a comparison with other non-adaptive methods.

This basis however is not, in general, optimal for class discrimination. The energy which best describes a given signal is not necessarily that which differentiates it from signals in other classes. A wavelet packet basis better suited for classification is more appropriately chosen using a *class separability index*, so as to provide a tiling that localizes the time-frequency regions containing the greatest class discrimination information. The promise of optimizing the segmentation of the time-frequency plane for classification provides the motivation for using wavelet packets over non-adaptive methods, such as the STFT or the wavelet transform. This subject is treated in detail in Chapter 3.

In addition to an appropriate tiling, it is often essential that *dimensionality reduction* be performed on the TFR, so that the information is sufficiently compact for presentation to a classifier. Although some dimensionality reduction *may* occur

as a result of transforming a signal into a TFR, in general it does not. Whereas the goal of the TFR is to localize and concentrate the signal's information, it is the role of dimensionality reduction to optimally *select* or *combine* the TFR coefficients. This issue was alluded to in Section 2.4.2, and is discussed in detail in the next section.

2.6 Dimensionality Reduction

Depending on the nature of a given classification problem (the raw data, the chosen features, and the chosen classifier) dimension reduction may be fundamental to successful classification performance. Dimensionality reduction is almost always essential when time-frequency representations are used as a feature basis; the descriptive ability of most TFRs comes at the expense of high dimensionality.

A schematic of the classification process is repeated in Figure 2.19, revised to include the notation used here to describe the data at each stage of processing. The input space $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^N$ and the response space $y \in \mathcal{Y} \subseteq \{y_1, \dots, y_K\}$ are as previously defined. The *feature extraction* stage is defined to be the initial transformation upon the measured signals, producing an “original” **feature set** which will be denoted $\mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^M$. This feature set \mathbf{v} is then subject to

dimensionality reduction, to yield a “reduced” feature set $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^L$ where $L < M$, which is more suitable for presentation to a classifier.

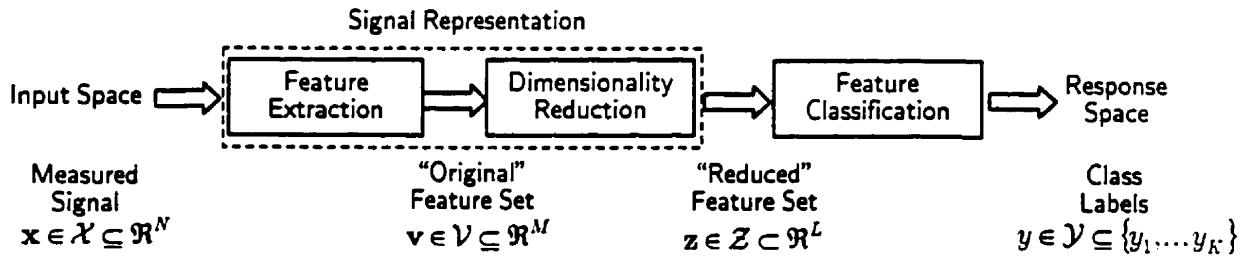


Figure 2.19 – The processing stages of classification: notation and dimensionality.

The principal motivation for dimensionality reduction is that it can help to alleviate the worst effects of the curse of dimensionality. The main goal is to ensure that as much of the relevant information as possible is preserved in as few dimensions as possible. A classifier with fewer inputs has fewer adaptive parameters to be determined, leading to a classifier with better generalization properties. Indeed, if we can perform sufficiently clever preprocessing, the classification task becomes trivial.

Dimensionality reduction strategies may be characterized as either *feature selection* or *feature projection*. These are discussed in the following sections.

2.6.1 Feature Selection

The feature selection approach attempts to reduce the number of variables by selecting the best *subset* of the original feature set, according to some criterion.

Feature selection necessarily consists of two components.

1. A criterion must be established by which it is possible to judge whether one subset of features is better than another.
2. A systematic procedure must be found for searching through candidate subsets of features.

Ideally, the selection criterion should be taken to be *the probability of misclassification*. In practice, evaluation of this criterion is generally too complex, and we have to resort to simpler criteria such as those based upon *class separability*. Similarly, in an ideal situation the search procedure should consist of an exhaustive search of all possible subsets. Exhaustive methods are often impractical due to computational complexity, and non-exhaustive searches and suboptimal (sequential) searches are often used in practice [Fukunaga90].

Class Separability Criteria

Consider, for simplicity, a two-class problem with an original feature set $\mathbf{v} \in \mathcal{V} \subseteq \mathbb{R}^M$. To simplify the problem further, assume that the candidate feature subsets consist of *single features*, instead of combinations of multiple features. This reduces the subset selection problem to that of applying the selection criterion to each feature individually, and *ranking* the M features in terms of their

discriminant power. In this case, the $L \leq M$ coefficients deemed most useful for classification may be chosen as the feature set.

Consider the feature matrices representing two different classes, designated as $[\mathbf{v}^{(p,1)} \ \mathbf{v}^{(p,2)} \ \dots \ \mathbf{v}^{(p,P)}]$ and $[\mathbf{v}^{(q,1)} \ \mathbf{v}^{(q,2)} \ \dots \ \mathbf{v}^{(q,Q)}]$, where p and q indicate the class, P and Q are the number of patterns in each class, and $\mathbf{v}^{(p,m)} = [v_1^{(p,m)} \ v_2^{(p,m)} \ \dots \ v_M^{(p,m)}]^T$ is the m^{th} pattern in class p . Expanded, these are:

$$\begin{bmatrix} v_1^{(p,1)} & v_1^{(p,2)} & \dots & v_1^{(p,P)} \\ v_2^{(p,1)} & v_2^{(p,2)} & \dots & v_2^{(p,P)} \\ \vdots & \vdots & \vdots & \vdots \\ v_M^{(p,1)} & v_M^{(p,2)} & \dots & v_M^{(p,P)} \end{bmatrix} \quad \begin{bmatrix} v_1^{(q,1)} & v_1^{(q,2)} & \dots & v_1^{(q,Q)} \\ v_2^{(q,1)} & v_2^{(q,2)} & \dots & v_2^{(q,Q)} \\ \vdots & \vdots & \vdots & \vdots \\ v_M^{(q,1)} & v_M^{(q,2)} & \dots & v_M^{(q,Q)} \end{bmatrix} \quad (2.35)$$

$v^{(p,2)}$ $v^{(q,2)}$

p_i } { q_i

When evaluating the discriminant power of each feature individually, the class separability criteria operate on single *rows* of these matrices. The discriminability of the i^{th} feature is determined by the i^{th} row of each matrix, which will be denoted as p_i and q_i for classes p and q , respectively.

We may define a *discriminant measure* for the i^{th} feature as $D(p_i, q_i)$. This measure should convey how separable p_i is from q_i . Class separability indices will

be characterized as belonging to one of three types, according to the nature of p_i and q_i .

Type I: *The data supplied to $\mathcal{D}(p_i, q_i)$ are the unprocessed features.*

If p_i and q_i represent the unprocessed features from the i^{th} row of the feature matrices, then we may apply Fisher's class separability index [Fisher50]:

$$\mathcal{D}(p_i, q_i) \doteq \frac{(\text{mean}(p_i) - \text{mean}(q_i))^2}{\text{var}(p_i) + \text{var}(q_i)} \quad (2.36)$$

Here, $\text{mean}_i(\cdot)$ and $\text{var}_i(\cdot)$ are operations to take the sample mean and variance across the i^{th} row of the feature matrices.

Type II: *The data supplied to $\mathcal{D}(p_i, q_i)$ represent the mean or the mean energy of the i^{th} feature.*

If the mean is computed across the i^{th} row of the feature matrix, p_i and q_i represent the average value of the i^{th} feature within each class. Similarly, if the mean square is computed across the i^{th} row, p_i and q_i represent the average energy of the i^{th} feature within each class (this is usually normalized by the total energy within the class). Similar measures may be taken using the moments or cumulants of the data. In this case, p_i and q_i are scalars, and several measures of class separability are applicable, including:

(i) **Relative Entropy:**

$$\mathcal{D}(p_i, q_i) \doteq p_i \log \frac{p_i}{q_i} \quad (2.37)$$

which measures the discrepancy of p_i from q_i [Kullback51], but is not symmetric with respect to p_i and q_i .

(ii) **Symmetric Relative Entropy:**

$$\mathcal{D}(p_i, q_i) \doteq p_i \log \frac{p_i}{q_i} + q_i \log \frac{q_i}{p_i} \quad (2.38)$$

which does yield symmetric activity amongst classes: clearly,

$$\mathcal{D}(p_i, q_i) = \mathcal{D}(q_i, p_i).$$

(iii) **Euclidean Distance:**

$$\mathcal{D}(p_i, q_i) \doteq \|p_i - q_i\| = (p_i - q_i)^2 \quad (2.39)$$

which is another asymmetric measure [Wantabe85].

Type III: *The data supplied to $\mathcal{D}(p_i, q_i)$ represent the probability density function of the i^{th} feature.*

This requires the estimation of the probability density functions of each feature. Sufficiently robust estimation of these densities usually involves kernel density estimators to reduce the estimation variance [Scott92]. Assume that an estimate of the continuous probability density function of the i^{th} feature in each class is available; these will be denoted $p_i(x)$ and $q_i(x)$. Two popular distance measures between probability density functions are:

(i) **Relative Entropy:**

$$\mathcal{D}(p_i, q_i) \doteq \int p_i(x) \log \frac{p_i(x)}{q_i(x)} dx. \quad (2.40)$$

(ii) **Hellinger Distance:** [Basseville89]

$$\mathcal{D}(p_i, q_i) \doteq \int (\sqrt{p_i(x)} - \sqrt{q_i(x)})^2 dx. \quad (2.41)$$

Whatever method is chosen, we may generalize the notation of the *discriminant measure* between measures p_i and q_i as $\mathcal{D}(p_i, q_i)$. If the discriminant measure is *additive* then the discriminant measure of subsets is easily evaluated⁶:

$$\mathcal{D}(\{p_i, p_j, p_k\}, \{q_i, q_j, q_k\}) = \mathcal{D}(p_i, q_i) + \mathcal{D}(p_j, q_j) + \mathcal{D}(p_k, q_k). \quad (2.42)$$

Indeed, we may define a set of M features as $\mathbf{p} = \{p_i\}_{i=1}^M$ (or $\mathbf{q} = \{q_i\}_{i=1}^M$), and evaluate the discriminability of this ensemble of features as:

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^M \mathcal{D}(p_i, q_i). \quad (2.43)$$

In general, it is necessary to measure discriminability amongst more than two classes. Let us abandon the class notation of p_i and q_i and let $p_i^{(k)}$, $k = 1, \dots, K$ represent the i^{th} feature of the k^{th} class, for each of K classes. Let $\mathbf{p}^{(k)} = \{p_i^{(k)}\}_{i=1}^M$ define a set of M features. To compute the discrepancy between the i^{th} feature in each class, we take the $\binom{K}{2}$ pairwise combinations of \mathcal{D} [Saito95]:

$$\mathcal{D}\left(\{p_i^{(k)}\}_{k=1}^K\right) \doteq \sum_{m=1}^{K-1} \sum_{n=m+1}^K \mathcal{D}(p_i^{(m)}, p_i^{(n)}). \quad (2.44)$$

Similarly, for an ensemble of features:

$$\mathcal{D}\left(\left\{\mathbf{p}^{(k)}\right\}_{k=1}^K\right) = \sum_{m=1}^{K-1} \sum_{n=m+1}^K \mathcal{D}(\mathbf{p}^{(m)}, \mathbf{p}^{(n)}). \quad (2.45)$$

Ideally, the process of feature selection should consider all possible subsets of the features. By evaluating each feature individually, one does not take into account the interactions between features. Interactions may be positive or negative: combinations of features can provide significant information that is not available in any of the individual features separately; conversely, two features that individually convey significant information may be highly correlated, and the pair may contribute little more information than either of the two.

As one considers subsets of multiple features however, the computational expense of evaluating the discriminant measures in multiple dimensions and of searching all combinations of feature subsets increases substantially. For this reason, suboptimal methods have been devised. These do not provide an exhaustive search of all possible feature subsets, but attempt to maximize the likelihood that a partial search will find a near-optimal solution. Among these methods are [Fukunaga90]:

- a) **The Add-On Method.** A measure of discriminability is evaluated for each feature, and the best feature is chosen as the first feature. Each of the remaining features is combined with the first feature to determine the best two-feature set. This process is iterated until the desired number of features have been determined.

⁶ Each of the class separability measures given above are additive.

- b) **The Knock-Out Method.** This method is similar to the Add-On method except that the feature set is pruned down to the desired number of features.

It should be noted that class separability measures do not invariably indicate classification performance. If we were able to use an ideal criterion of misclassification rate, we would expect that, as the number of features retained was reduced, the generalization performance of the chosen classifier would improve (as a consequence of the curse of dimensionality). This would continue until some optimal feature set size was reached, and that if fewer features were retained the performance would degrade. Most class separability indices are incapable of modeling this phenomenon; the criteria will always improve as more features are added [Bishop96]. Class separability criteria are useful, however, for comparing different feature sets of equal size.

In general, feature selection methods utilize class membership in the determination of discriminant power. The original identity of the features is maintained, which is useful in interpreting the utility of individual features. The main disadvantage of feature selection methods is that some of the original features and therefore, the information that they convey, must be discarded completely. This is a problem if class separability information is broadly dispersed throughout the original feature set. For this reason feature selection methods do not work well on high-dimensional and highly redundant data, which is often the case with unprocessed measured waveforms. Unless the data are well-behaved, information in high-dimensional spaces tends to be dispersed, and the likelihood that a few features contain the majority of class separability information is small. This implies that feature selection methods are best applied to original feature sets that suitably localize or concentrate the invariant characteristics of the input space.

2.6.2 Feature Projection

As opposed to feature selection, which seeks to select the best *subset* of the original feature space for class separability, the goal of feature projection methods is to determine the best *combination* of the original features to form a (generally smaller) feature set. For classification, the projection $\mathcal{F} \rightarrow \mathcal{Z}$ should map the data into separate clusters, one per class, facilitating the classification task. If the map $\mathcal{F} \rightarrow \mathcal{Z}$ is *linear*, the optimization task is simply to find the coefficients of the linear function so as to minimize or maximize the chosen criterion. If this criterion is simple, the well-developed techniques of linear algebra may be used. For complex criteria, it is usually necessary to apply iterative techniques to determine the mapping coefficients.

2.6.2.1 Principal Components Analysis

Consider first the task of feature projection for accuracy of signal representation – how accurately can a set of features represent a distribution of data? If a small set of features may be used to reconstruct the signal accurately, then the features may be said to be effective. Clearly, this has direct application in signal compression. The most intuitive criterion for signal compression is the *minimum mean-square error* (MMSE). Principal components analysis (PCA) provides a linear map which minimizes the MMSE criterion⁷. PCA's effectiveness in pattern recognition is due to its ability to eliminate linear dependencies and uncorrelated noise in the data [Huber85].

⁷ PCA is known synonymously as the *Karhunen-Loéve transform* and the *singular value decomposition*.

Consider a dataset of P feature vectors $\mathfrak{I} = [\mathbf{v}^{(1)}, \dots, \mathbf{v}^{(P)}]$, where any given feature vector in the set is $\mathbf{v}^{(p)} = [v_1^{(p)} \dots v_M^{(p)}]^T$. The goal of PCA is to map each original feature vector $\mathbf{v}^{(p)} \in \mathcal{V} \subseteq \mathbb{R}^M$ onto a reduced feature vector $\mathbf{z}^{(p)} = [z_1^{(p)} \dots z_L^{(p)}]^T \in \mathcal{Z} \subseteq \mathbb{R}^L$, $L < M$, subject to a MMSE criterion. Specifically, a given feature vector $\mathbf{v}^{(p)}$ can be represented without error by the summation of M orthonormal vectors as

$$\mathbf{v}^{(p)} = \sum_{i=1}^M z_i^{(p)} \mathbf{u}_i. \quad (2.46)$$

Explicit expressions for the coefficients $z_i^{(p)}$ can be found by

$$z_i^{(p)} = \mathbf{u}_i^T \mathbf{v}^{(p)}. \quad (2.47)$$

Suppose that we retain only a subset $L < M$ of the basis vectors \mathbf{u}_i , so that we only use L coefficients $z_i^{(p)}$. The remaining coefficients will be replaced by coefficients b_i so that each vector $\mathbf{v}^{(p)}$ is approximated by an expression of the form⁸:

$$\tilde{\mathbf{v}}^{(p)} = \sum_{i=1}^L z_i^{(p)} \mathbf{u}_i + \sum_{i=L+1}^M b_i \mathbf{u}_i. \quad (2.48)$$

This represents a form of dimensionality reduction, since the original feature vector $\mathbf{v}^{(p)}$ is approximated by a new vector $\mathbf{z}^{(p)}$, which has $L < M$ degrees of freedom. Now consider the whole dataset of P vectors \mathfrak{I} . We wish to choose the basis vectors \mathbf{u}_i and the coefficients b_i such that the approximation gives the best approximation to each original vector, on average, for the whole dataset. The sum of the squared errors over the whole dataset is

⁸ Note that the coefficients b_i are independent of the pattern p ; these provide a bias term in the estimate.

$$E_L \doteq \frac{1}{2} \sum_{p=1}^P \left\| \mathbf{v}^{(p)} - \bar{\mathbf{v}}^{(p)} \right\|^2 = \frac{1}{2} \sum_{p=1}^P \sum_{i=L+1}^M \left(z_i^{(p)} - b_i \right)^2. \quad (2.49)$$

To minimize the error with respect to the b_i , we set the derivative of E_L to zero, with the result:

$$b_i = \frac{1}{P} \sum_{p=1}^P z_i^{(p)} = \mathbf{u}_i^T \bar{\mathbf{v}} \quad (2.50)$$

where we have defined the mean vector $\bar{\mathbf{v}} = \sum_{p=1}^P \mathbf{v}^{(p)}$. Using these results, we can write the sum-of-squares error as

$$E_L = \frac{1}{2} \sum_{p=1}^P \sum_{i=L+1}^M \left\{ \mathbf{u}_i^T (\mathbf{v}^{(p)} - \bar{\mathbf{v}}) \right\}^2 = \frac{1}{2} \sum_{i=L+1}^M \mathbf{u}_i^T \Sigma_v \mathbf{u}_i, \quad (2.51)$$

where Σ_v is the covariance matrix of the set of vectors \mathbf{v} :

$$\Sigma_v \doteq \sum_{p=1}^P (\mathbf{v}^{(p)} - \bar{\mathbf{v}})(\mathbf{v}^{(p)} - \bar{\mathbf{v}})^T \quad (2.52)$$

There now remains the task of minimizing E_L with respect to the basis vectors \mathbf{u}_i . [Bishop96] has shown that the minimum occurs when the basis vectors satisfy

$$\sum_v \mathbf{u}_i = \lambda_i \mathbf{u}_i \quad (2.53)$$

so that the \mathbf{u}_i are the eigenvectors and the λ_i are the eigenvalues of Σ_v . Each of the eigenvectors \mathbf{u}_i is called a *principal component*.

In practice, the algorithm proceeds by computing the mean of the dataset $\bar{\mathbf{v}} = \sum_{p=1}^P \mathbf{v}^{(p)}$ and subtracting off this mean. Then the covariance matrix and the eigenvalues are found. The eigenvectors corresponding to the L largest eigenvalues are retained and the original feature vectors \mathfrak{I} are projected onto the eigenvectors to yield an L -dimensional feature set.

In the context of pattern recognition, the coefficients z_1, \dots, z_L are viewed as the feature set. It may be regarded as a form of *unsupervised* learning, as class membership is not used in the MMSE criterion. This feature space has some desirable attributes [Fukunaga90]:

1. The effectiveness of each feature, in terms of representing \mathbf{v} , is determined by its corresponding eigenvalue. The features are *ranked* and those with the largest eigenvalues are kept as a feature set.
2. The feature values are mutually uncorrelated.

In some cases however, the best criteria for signal compression are not the same as for signal classification. As an example, consider a theoretical distribution of height and weight for males and females, depicted in Figure 2.20. Since these two variables (features) are highly correlated (a taller person is usually heavier), the shape of the distribution for each gender is oblong. PCA would yield projections along the two principal axes \mathbf{u}_1 and \mathbf{u}_2 , yielding the features z_1 and z_2 respectively. According to the MMSE criterion, the principal axis \mathbf{u}_1 is a better vector than \mathbf{u}_2 to represent the distribution. Note however, that if the distributions are mapped onto \mathbf{u}_1 , the marginal density functions overlap. On the other hand, if the distributions are mapped onto \mathbf{u}_2 , the marginal densities exhibit very little overlap, indicating better class separability. Therefore, for classification purposes, \mathbf{u}_2 is a better feature extractor than \mathbf{u}_1 , preserving more of the class information.

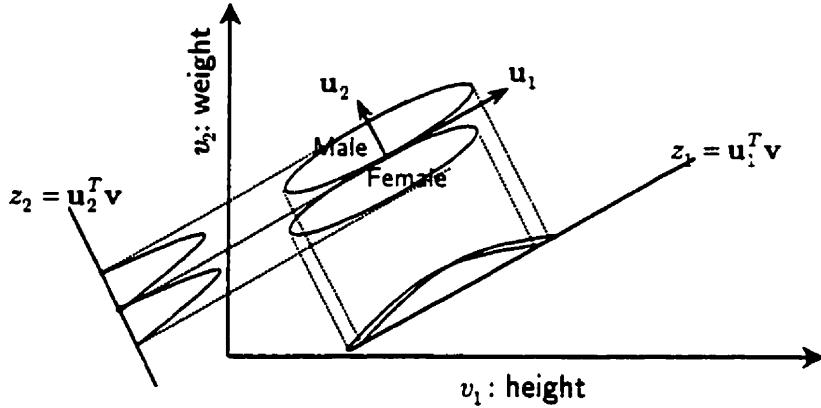


Figure 2.20 – The difference between feature extraction for signal compression and signal classification.

This problem is somewhat contrived, however; PCA does tend to cluster many datasets effectively. The uncorrelated feature set is devoid of any linear dependencies, and PCA acts as a variation reducing technique, relegating most of the random noise to the trailing components and embedding the systematic structure into the leading ones. For these reasons, PCA often performs very well as a dimensionality reduction technique when classifying datasets of physical origin.

2.6.2.2 Projection Pursuit

Separability criteria are invariably more complex than the simple MMSE criteria used in signal compression. *Projection pursuit* (PP) was first proposed by Kruskal [Kruskal69], and was first implemented (and named) by Friedman and Tukey [Friedman74]. The original purpose of PP was to pick the most “interesting” low-dimensional projections of high-dimensional data clouds automatically by

optimizing a specified objective function or *projection index*. Depending upon the nature of the projection index, the PP algorithm may be tailored for compression, classification, regression or density estimation. Huber [Huber85] describes a number of projection indices – with relative merits – suitable for class discrimination.

The most promising feature of PP is that it is one of the very few methods of processing multivariate data capable of “bypassing” the curse of dimensionality. PP avoids this problem by working in low-dimensional linear projections. PP can be applied to an M -dimensional data cloud to determine the best L -dimensional projection, subject to the projection index⁹. In this case, a direct application of PP would require the optimization of an L -dimensional projection index, which is likely to be extremely computationally intense if $L > 2$ and M is large. To make the algorithm computationally manageable, one can find the sequence of best one-dimensional projections by optimizing the projection index, removing the structure that makes this projection interesting, and re-iterating.

It seems then, that PP with a projection index chosen to measure the discriminant power of projections is an attractive approach for dimensionality reduction. There are some problems with PP, however:

1. A sequential acquisition of the best one-dimensional projections may be too “greedy”. Important multiple-dimensional structures may not be recognized as being important by the one-dimensional search algorithm.

⁹ As a matter of interest, it can be shown that given an appropriate projection index, many classical multivariable data analysis techniques, including PCA and LDA, are special cases of PP.

2. Although the projection index can guide the algorithm, it is an exploratory method and therefore it is difficult to know whether an optimal solution has been reached. The ability of the algorithm to cluster the data is very sensitive to the choice of projection index.
3. Classical approaches have used nonlinear programming techniques that search the M -dimensional feature space for “interesting” projections, but these methods require enormous computational expense as M increases beyond ten dimensions [Friedman74]. Neural network approaches have been proposed that alleviate the computational complexity (see the next section), but they are equally vulnerable to the effects of local minima and are sensitive to the choice of projection index [Intrator92].

2.6.2.3 Neural Networks

In many pattern recognition problems there are important features that are *not* linear functions of the original measurements, but are highly nonlinear functions. As yet, there are no systematic methods for generating nonlinear mapping functions. Multilayer neural networks however, have been shown to have some success in providing nonlinear dimensionality reduction.

Consider first a multilayer perceptron network, shown in Figure 2.21, having M inputs, M outputs, and L hidden layer units, with $L < M$. The targets used to train the network are simply the input vectors themselves, so that the network is attempting to map each input vector onto itself, using a MMSE cost function. Due to the reduced number of units in the hidden layer, a perfect reconstruction of all input vectors is not possible.

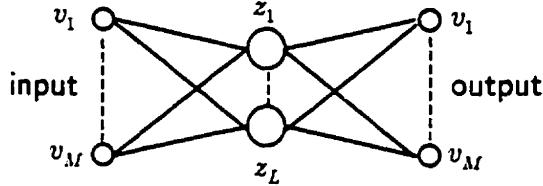


Figure 2.21 – An auto-associative MLP having two layers of weights. The network is trained by mapping the input vector onto itself via a MMSE cost function. Even with nonlinear activation functions in the hidden layer, such a network is equivalent to a linear PCA. Biases have been omitted for clarity.

Such a network is said to form an *auto-associative* mapping. If the hidden layer units have linear activation functions, it can be shown that the MMSE has a global minimum, and that at this minimum the network performs a projection onto the L -dimensional subspace which is spanned by the first L principal components of the data [Baldi89]. The first layer weights therefore specify the projection onto the principal subspace. It might be thought that the limitations of linear dimensionality reduction might be overcome by using nonlinear activation functions for the hidden units in this network. It has been shown however, that such nonlinearities make no difference [Bourlard88]. In a similar manner, PCA has been implemented using a class of neural networks that use the Hebbian learning rule, with the weights converging to orthonormal vectors along the principal component directions [Oja82][Sanger89].

There is no practical benefit, however, in using two-layer MLP networks or Hebbian networks to perform dimensionality reduction. Standard techniques for PCA (based on the singular value decomposition) are guaranteed to give the correct solution in finite time, and also generate an ordered set of eigenvalues with corresponding eigenvectors.

Some benefit is gained however, if additional layers are permitted in a MLP network. A four-layer MLP network with sigmoidal nonlinear activation functions in the first and third layers has been demonstrated to be capable of performing a nonlinear PCA [Kramer91]. The minimization of the network error however, is a nonlinear optimization problem, requiring computationally intensive techniques. Also, there is a risk of finding suboptimal local minima.

Another neural network solution of providing nonlinear projections are *self-organizing maps*, such as those proposed by Kohonen [Kohonen90]. Although these maps have interesting nonlinear clustering capabilities, training again involves the same problems associated with nonlinear optimization. Specifically, it is difficult to know how well the network has done; that is, if the network has converged to a global minimum.

Intrator [Intrator92] has described a neural network which appears to perform a form of projection pursuit. This network is based upon an artificial model of neurons found in the visual cortex, known as BCM neurons [Bienenstock82]. Although this neural network does eliminate some of the computational complexity of standard algorithms, it is not clear that it is actually performing a projection pursuit. The network's performance is sensitive to the choice of objective function and its parameters, and to the number of output neurons (the dimension of the reduced feature set) [Gallant93]. The unsupervised nature of the BCM training algorithm makes it difficult to know whether local minima have been encountered. Consequently, a robust stopping criterion has been elusive.

2.6.3 Summary: Dimensionality Reduction

Having explained the relative merits of feature selection and feature projection, it is left to discover which is best suited to a given problem. It is the nature of the measured signals and of the chosen original feature set that will determine the efficacy of dimensionality reduction methods. The measured signals here are transient MES patterns, and the feature sets of primary interest are TFRs of various flavors. It is fairly certain that the success of a chosen TFR as a classification feature set depends upon the proper design of a dimensionality reduction method. Dimensionality reduction, therefore, will play a key role in the investigation of signal representation for classification in Chapter 4.

2.7 Summary

The intent of this chapter has been to provide a global perspective on the problem of signal classification, and to provide some insight into the issues specific to transient signal representation. The aspects of the classification problem are perhaps seen with greater clarity by partitioning the task into the stages of signal representation and classifier design. Following an overview of some classifiers and their relative merits, it has been emphasized that signal representation is the element crucial to classification performance.

The process of signal representation consists of a *feature extraction* stage (which generates an “*original feature set*”) and, in general, a *dimensionality reduction* stage (which pares the data down to a “*reduced feature set*” suitable for presentation to a classifier). It is proposed that the time-frequency domain provides a robust and versatile framework for feature extraction, with the expectation that this two-dimensional map concentrates discriminant information more effectively than one-dimensional alternatives in either time or frequency.

The properties of several time-frequency energy representations (linear and quadratic) were discussed. The issues that are important to classification of transient myoelectric signals are (i) the ability of the transform to localize the information needed to discriminate amongst classes and (ii) the computational complexity of the algorithm. The complexity of the linear methods discussed here are comparable: the STFT is $O(N \log N)$, the wavelet transform is $O(N)$ and wavelet packet methods require $O(N \log N)$. Quadratic methods are more

expensive: they are generally $O(N^2)$, plus extra processing to compute the analytic signal and to implement a smoothing kernel, if needed. For this reason, they will not be considered as a basis for feature extraction.

For linear methods, the issue of interest is the manner in which they tile the time-frequency plane. Linear methods are necessarily constrained by the Heisenberg bound; the difference lies in the geometry of the time-frequency cells. The STFT provides an adequate representation if the information is well contained by a rectangular grid of fixed aspect ratio. This is often not the case with signal of physical origin, including the transient MES patterns of interest here. The WT provides a dyadic tiling that is somewhat more appropriate, in a general sense. The WPT, however, allows the tiling to be adaptively chosen. A properly chosen cost function can specify the optimal WPT basis for energy concentration (signal compression) or class separability (signal classification) for a given set of signals. This adaptive basis may be expected to provide superior localization, thereby conveying a maximum amount of information in the fewest number of cells. In light of the promise shown by the WPT, a significant portion of Chapter 3 is dedicated to the application of the WPT as a basis for feature extraction.

Chapter 3

Wavelet-Based Feature Extraction

3.1 Introduction

At the present time, the scope of wavelet research is growing rapidly in many directions, and it is very difficult to rigorously define a wavelet. One can choose between smooth wavelets, compactly supported wavelets, wavelets with simple mathematical expressions, wavelets with simple associated filters, etc. We may, however, develop some generalities that form the foundation of wavelet theory. Denote a wavelet as $\psi_\lambda(x)$, where x belongs to some domain \mathcal{X} , and λ is an index belonging to a set Λ . We may refer to $\Psi = \{\psi_\lambda | \lambda \in \Lambda\}$ as a *wavelet basis*. We may make three statements that provide a general characterization of wavelet bases, and which deliver some insight into their signal processing capabilities:

1. Wavelets are building blocks for general functions. Any general function $f \in L^2(\mathbb{R})$ may be expressed as an infinite series of wavelets¹. That is, a set of coefficients must exist such that

$$f = \sum_{\lambda} c_{\lambda} \psi_{\lambda}. \quad (3.1)$$

2. Wavelets have time-frequency localization. Locality in time means that most of the energy of the wavelet is restricted to a finite interval. If the function is identically zero outside a given interval, it is termed *compactly supported*. Frequency localization means that the Fourier transform of the wavelet is localized (*i.e.* it is bandlimited). The Heisenberg uncertainty principle puts a lower bound on the product of time and frequency variance. Instead of a fixed time and frequency resolution, as given by the short-time Fourier transform, a wavelet analysis varies the time-frequency aspect ratio, producing good frequency localization at low frequencies (long time windows), and good time localization at high frequencies (short time windows). This produces a segmentation or *tiling* of the time-frequency plane that is appropriate for most physical signals, especially those of a transient nature.

3. Wavelets have fast transforms. In many applications, it is important that the transforms be easy to implement. In the case of real-time processing, as is the problem addressed here, it is essential. Fast wavelet transforms are obtained through *multiresolution analysis*; a pyramid algorithm with its origins in image processing that was adapted for wavelet analysis by Stephane Mallat

¹ The space $L^2(\mathbb{R})$ is the space of square-integrable functions. A function f is in $L^2(\mathbb{R})$ if $\int f^2 < \infty$.

and Yves Meyer [Meyer93]. The fast wavelet transform uses a series of linear filters – lowpass and highpass – to decompose the signal into low and high-frequency components. The algorithm also combines these filters with *downsampling* operations, that is, steps that decimate the signal at each stage, halving the data each time. This feature accounts for the algorithm’s speed, because the *downsampling* reduces the computations at each iteration geometrically – at j iterations, the number of samples being manipulated shrinks by 2^j . This yields very efficient algorithms: most N -point wavelet transforms have complexity on the order of $O(N)$, whereas a Fourier transform is of the order $O(N \log N)$ [Sweldons96][Vetterli95]. If the transform’s complexity is CN (where C is a constant), then C depends on the wavelet chosen [Bruce96]. If C is small, then computing the wavelet transform requires about the same effort as trivial tasks such as copying or rescaling a signal. Wavelets involving only a few terms subtend great efficiency (a small value of C), such as those developed by Ingrid Daubechies [Daubechies92].

This chapter is intended to provide a mathematical basis for the wavelet transform (WT), the wavelet packet transform (WPT), and the cosine packet transform (CPT), and to describe how they may be optimized for classification. Section 3.2 describes wavelet theory using subspace mathematics and, for the sake of clarity, in the language of signal processing as well. The relationship between wavelet bases and their time-frequency localization is developed, so that a region of a TFR can be identified with a certain basis function. Section 3.3 introduces the WPT and the CPT, which are generalizations of the wavelet transform. The nature of this generalization is such that they can adaptively partition the time-frequency plane; the WPT optimizes a criterion with respect to frequency, and the CPT with

respect to time. Section 3.4 discusses this optimization problem, which has been termed *best basis selection*. Best basis selection is first described in the context of signal compression and then, a modified algorithm for signal classification is presented. The chapter concludes with a simple example that demonstrates wavelet-based feature extraction for classification.

3.2 Wavelet Bases

This section provides a brief introduction to wavelet theory. This is by no means meant to be an exhaustive treatment of wavelet mathematics, but simply to provide a reference for the nomenclature used in later sections. Several excellent texts offer a complete background for the interested reader [Vetterli95][Meyer93] [Kaiser95][Daubechies92].

3.2.1 The Continuous Wavelet Transform

As described in Chapter 2, the continuous wavelet transform (CWT) is defined to be

$$\text{CWT}_x(\tau, a) \doteq \int x(t) \psi_{a,\tau}^*(t) dt \quad (3.2)$$

where

$$\psi_{a,\tau}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-\tau}{a}\right), \quad (3.3)$$

where $\psi(t)$ is the *mother wavelet*. The mother wavelet has the property that the set $\{\psi_{a,\tau}(t)\}_{a,\tau \in \mathbb{Z}}$ forms an orthonormal basis in $L^2(\mathbb{R})$. This implies that the mother wavelet can, in turn, generate any function in $L^2(\mathbb{R})$.

Figure 3.1 demonstrates the nature of a symmetric wavelet at various scales and translations. Note that, at small scales, a temporally localized analysis is done; as scale increases, the breadth of the wavelet function increases, thereby analyzing with less time resolution but greater frequency resolution. Notice, as well, that the wavelet functions are *bandpass* in nature, thus partitioning the frequency axis. In fact, a fundamental property of wavelet functions is that

$$c = \frac{\Delta f}{f}, \quad (3.4)$$

where Δf is a measure of the bandwidth, f is the centre frequency of the passband, and c is a constant. The wavelet functions may therefore be viewed as a bank of analysis filters with a constant relative passband (a "constant-Q" analysis).

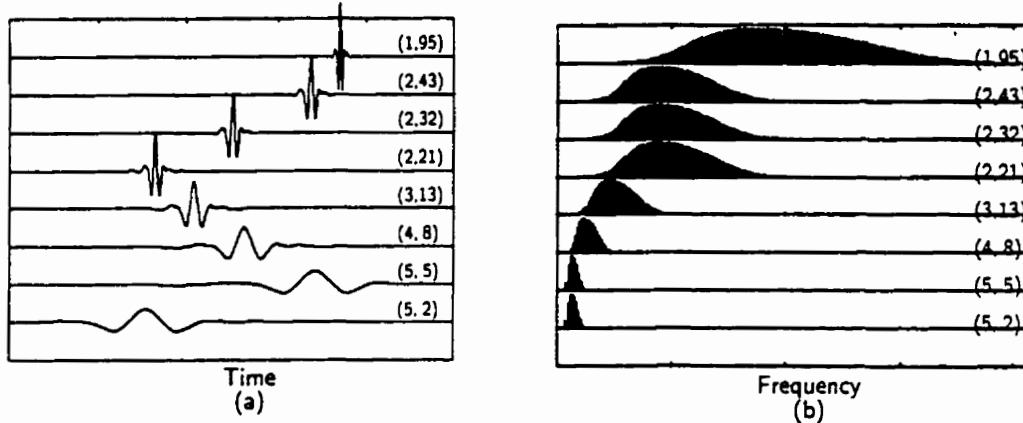


Figure 3.1 – Some symmetric wavelets at various scales and locations.
The couplet (a, τ) marking each waveform denotes scale and shift, respectively.
Figure (a) shows the time domain, and (b) the frequency domain of each wavelet.

Recall the discussion of time-frequency resolution in the previous chapter. For the STFT we have a fixed time resolution Δt and frequency resolution Δf constrained by the Heisenberg inequality $\Delta t \cdot \Delta f \geq \frac{1}{4\pi}$, producing a regular grid in the time-frequency plane. Consider now the case of a wavelet analysis. If the time resolution and frequency resolution of the mother wavelet $\psi(t)$ are Δt and Δf , respectively, it can be shown that for the scaled-shifted wavelet $\psi_{a,r}(t)$, the time and frequency resolutions are

$$\Delta t' = |a| \Delta t \quad \text{and} \quad \Delta f' = \frac{1}{|a|} \Delta f. \quad (3.5)$$

It is clear that the aspect ratio of a cell in the time-frequency plane depends upon the scale a . Thus, at low scale (high frequency) the Heisenberg box is "narrow and

tall”, while at high scale values (low frequency), the time-frequency cell is “wide and short².“ Overall, the time-frequency localization of each wavelet function is still subject to the Heisenberg bound:

$$\Delta t' \cdot \Delta f' = \Delta t \cdot \Delta f \geq \frac{1}{4\pi}. \quad (3.6)$$

The classic example of continuous-time wavelet analysis uses a windowed complex exponential as the mother wavelet. This is the *Morlet wavelet* [Grossmann84] given by

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-j\omega_0 t} e^{-t^2/2}, \quad (3.7)$$

where ω_0 is the centre frequency; a value of $\omega_0 = 5.336$ is commonly used to provide a desirable wavelet shape [Vetterli95].

3.2.2 The Discrete Wavelet Transform

To analyze discrete-time signals, it is convenient to take special values for a and τ in defining this basis: if $a = 2^j$ and $\tau = n \cdot 2^j$ (where j and n are integers) then, via translations and dilations:

$$\{\psi_{j,n}(t)\}_{j,n \in \mathbb{Z}} \doteq \{2^{-j/2} \psi(2^{-j}t - n)\} \quad (3.8)$$

forms a sparse orthonormal basis of $L^2(\mathbb{R})$ [Meyer93]. This means that the wavelet basis induces an orthogonal decomposition of any function in $L^2(\mathbb{R})$:

$$L^2(\mathbb{R}) = \bigoplus_j \Omega_{j,1} \quad (3.9)$$

² Assuming that time is represented along the horizontal axis and frequency along the vertical axis.

where $\Omega_{j,1}$ is the subspace spanned by $\{\psi_{j,n}\}_{n \in \mathbb{Z}}$. Thus, a complete description of the original signal is available from a direct sum of orthogonal subspaces. The subscript “1” is used to differentiate $\Omega_{j,1}$ from its dual space $\Omega_{j,0}$, which is defined as follows. A wavelet function is always associated with a companion: the *scaling function*, $\varphi(t)$, which is also sometimes called the *father wavelet*. Like the wavelet function, the scaling function,

$$\{\varphi_{j,n}(t)\}_{j,n \in \mathbb{Z}} = \{2^{-j/2} \varphi(2^{-j}t - n)\}_{j,n \in \mathbb{Z}} \quad (3.10)$$

forms a sparse orthonormal basis of $L^2(\mathbb{R})$. The scaling function induces a chain of nested subspaces

$$\Omega_{J,0} \subset \Omega_{J-1,0} \subset \dots \subset \Omega_{1,0} \subset \Omega_{0,0} \quad (3.11)$$

where $\Omega_{j,0}$ is the subspace spanned by $\{\varphi_{j,n}\}_{n \in \mathbb{Z}}$.

What does this really mean? The nature of the scaling function is that a *projection* of the original signal $x(t)$ onto the space $\Omega_{j,0}$ is a *lowpass* operation.

Specifically, the projection of $x(t)$ onto $\Omega_{j,0}$ is an *approximation* at scale $a = 2^j$:

$$A_j[n] = \langle x(t), \varphi_{j,n}(t) \rangle \quad (3.12)$$

We define $\Omega_{0,0}$ (scale $a = 2^0 = 1$) to be the space of the original signal $x(t)$; that is, $A_0[n] = x[n] \equiv x(nT_s)$. Thus, $\Omega_{J,0} \subset \Omega_{J-1,0} \subset \dots \subset \Omega_{1,0} \subset \Omega_{0,0}$ is actually a sequence of successively coarser approximations of $x(t)$ as scale ranges from 0 to J . The subspaces subtended by the wavelet and the scaling functions are related such that:

$$\Omega_{j,0} = \Omega_{j+1,0} \oplus \Omega_{j+1,1} \quad \text{for } j = 0, 1, \dots, J \quad (3.13)$$

meaning that $\Omega_{j,1}$ contains the *detail* needed to go from a coarser to a finer level of approximation. The detail component of $x(t)$ at scale $a = 2^j$ is:

$$D_j[n] = \langle x(t), \psi_{j,n}(t) \rangle. \quad (3.14)$$

This is a bandpass operation. Therefore, the wavelet transform may be viewed as a way to represent $\Omega_{0,0}$ as a direct sum of mutually orthogonal subspaces:

$$\Omega_{0,0} = \left(\bigoplus_{j=1}^J \Omega_{j,1} \right) \oplus \Omega_{J,0}. \quad (3.15)$$

This concept is perhaps more clearly illustrated in Figure 3.2. The wavelet transform processes a signal by decomposing it into successive approximation $A_j[n] \in \Omega_{j,0}$ and detail $D_j[n] \in \Omega_{j,1}$ signals. The approximation signal is re-sampled at each stage, and the detail coefficients are kept. The aim of the analysis is to arrive, starting from the original sampled signal $x[n] = A_0[n]$, at a decomposition:

$$\{ D_1[n], D_2[n], \dots, D_J[n], A_J[n] \}. \quad (3.16)$$

This is the wavelet transform: for a decomposition into J scales, the transform coefficients consist of J scales of detail coefficients and, at the J^{th} scale, the lowest-level approximation signal.

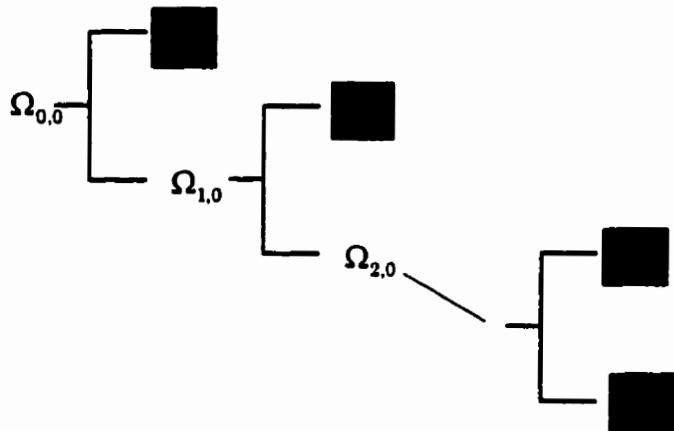


Figure 3.2 – Subband decomposition analogy of the wavelet transform. The symbols in gray represent those subspaces kept intact by the wavelet transform.

3.2.3 Multiscale Filter Banks

From an electrical engineering perspective, it is instructive to repeat the theory of wavelet decomposition in the language of signal processing. Let the map $\Omega_{j,0} \rightarrow \Omega_{j+1,0}$ be represented by the operator H , and the map $\Omega_{j,0} \rightarrow \Omega_{j+1,1}$ be represented by G . This is illustrated in Figure 3.3 as an operation on a N -sample signal $\mathbf{x} = \{x[n]\}_{n=0}^{N-1} \in \Omega_{0,0}$.

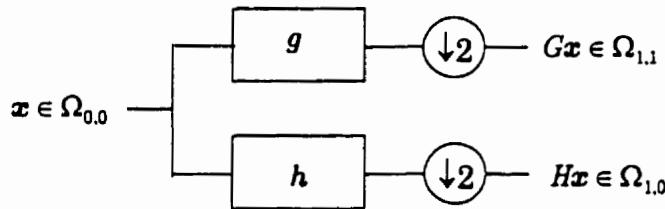


Figure 3.3 – Projection operators H and G .

As the figure implies, the operators consist of two separate stages: $\mathbf{g} = \{g[n]\}_{n=0}^{L-1}$ and $\mathbf{h} = \{h[n]\}_{n=0}^{L-1}$ are highpass and lowpass filters, respectively, and the $\downarrow 2$ operator implies a decimation by 2. This may be written as:

$$(H\mathbf{x})_n = \sum_{k=0}^{L-1} h[k] x[2n - k], \quad (G\mathbf{x})_n = \sum_{k=0}^{L-1} g[k] x[2n - k], \quad (3.17)$$

for $n = 0, 1, \dots, N-1$. This makes intuitive sense, that the approximation $\Omega_{j,0} \rightarrow \Omega_{j+1,0}$ should be obtained via lowpass filtering, and the detail $\Omega_{j,0} \rightarrow \Omega_{j+1,1}$ by highpass filtering. If $\mathbf{x} = \{x[n]\}_{n=0}^{N-1} \in \Re^N$ is a vector to be analyzed, the operators transform the vector \mathbf{x} into two subsequences $G\mathbf{x}$ and $H\mathbf{x}$, of length $N/2$. Next, the same operations are applied to the vector of the lower frequency band $H\mathbf{x}$ to obtain $H^2\mathbf{x}$ and $GH\mathbf{x}$ of lengths $N/4$. If the process is repeated $J \leq \log_2 N$ times, the wavelet decomposition may be then written as:

$$\{G\mathbf{x}, GH\mathbf{x}, GH^2\mathbf{x}, \dots, GH^J\mathbf{x}, H^{J+1}\mathbf{x}\}, \quad (3.18)$$

of length N . This is demonstrated diagrammatically in Figure 3.4.

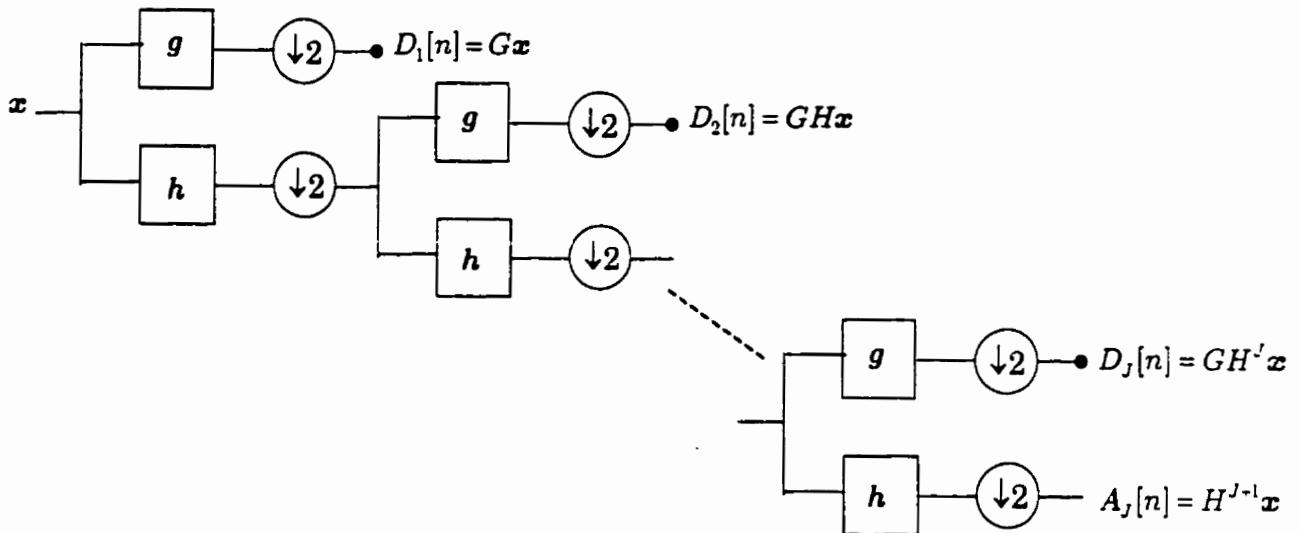


Figure 3.4 – The subband coding analogy of the DWT.

The wavelet transform thus analyzes the data by partitioning its frequency content dyadically finer and finer toward the low frequency region (and coarser and coarser in the time domain).

The issue to be addressed now is, how are the filters \mathbf{h} and \mathbf{g} determined? The operators H and G are called *perfect reconstruction* or *quadrature mirror filters* (QMFs) if they satisfy the following orthogonality conditions:

$$HG^* = GH^* = 0 \quad \text{and} \quad HH^* = GG^* = I, \quad (3.19)$$

where $*$ is the adjoint operator and I is the identity matrix. These conditions impose some restrictions on the filter coefficients of \mathbf{h} and \mathbf{g} . Let m_0 and m_1 be the bounded periodic functions defined by

$$m_0(\xi) = \sum_{n=0}^{L-1} h[n]e^{jn\xi}, \quad m_1(\xi) = \sum_{n=0}^{L-1} g[n]e^{jn\xi}. \quad (3.20)$$

Daubechies proved [Daubechies88] that H and G are QMFs *iff* the following matrix is unitary for all $\xi \in \Re$:

$$\begin{pmatrix} m_0(\xi) & m_0(\xi + \pi) \\ m_1(\xi) & m_1(\xi + \pi) \end{pmatrix}. \quad (3.21)$$

Various design criteria (concerning regularity, symmetry, *etc.*) on the lowpass filter coefficients h can be found in [Daubechies92]. Once the h are fixed, we can specify the QMFs by setting $g[n] = (-1)^n h[L-1-n]$.

The QMFs are related to the wavelet and scale functions by [Meyer93]:

$$h[n] = \langle \varphi(t), \sqrt{2}\varphi(2t - n) \rangle \quad (3.22)$$

and

$$g[n] = \langle \psi(t), \sqrt{2}\psi(2t - n) \rangle. \quad (3.23)$$

If H and G are QMFs, then the perfect reconstruction property allows exact reconstruction of the original signal from the wavelet transform coefficients. Since the wavelet bases are orthogonal, reconstruction takes the form of *upsampling* (inserting zeros at every other sample) and filtering by the QMFs; exactly the reverse operation of the subband decomposition scheme³.

3.2.4 A Simple Example

After a highly mathematical treatment of the wavelet transform, an illustrative example might be helpful. Consider the linear chirp signal ($N = 256$) which was

analyzed in the previous chapter. This signal may be decomposed into its detail and approximation signals:

| Signal | $x(t)$ | D_1 | D_2 | D_3 | D_4 | D_5 | D_6 | D_7 | D_8 | A_8 |
|--------|--------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| Length | 256 | 128 | 64 | 32 | 16 | 8 | 4 | 2 | 1 | 1 |

A Daubechies-4 wavelet has been chosen for this analysis [Daubechies92]; the wavelet and scaling functions are shown in Figure 3.5. The issues regarding wavelet properties and wavelet selection are discussed in the next section.

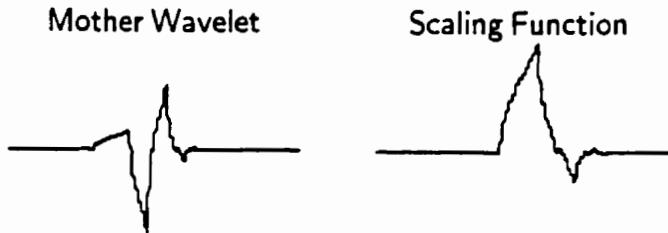


Figure 3.5 – The Daubechies-4 wavelet and scaling functions

Although a decomposition is usually carried out to its maximal depth $J = \log_2 N$, the decomposition may terminate at any level. This means that the process of dividing the frequency axis into finer and finer segments toward the low frequency range stops, and the final detail and approximation signals are of length greater than one. This may be desirable if the subdivision of bands beyond a certain scale does not yield subbands with a significant energy component.

³ For orthogonal wavelet bases, the filters \mathbf{g} and \mathbf{h} are used for both decomposition and synthesis. An alternative scheme, referred to as biorthogonal wavelet bases, allows perfect reconstruction with synthesis filters that are different than the analysis filters [Cohen90].

In Figure 3.6, we see the results of a wavelet decomposition of the chirp signal $x(t)$. The decomposition was terminated at a depth of scale 3 for clarity of presentation.

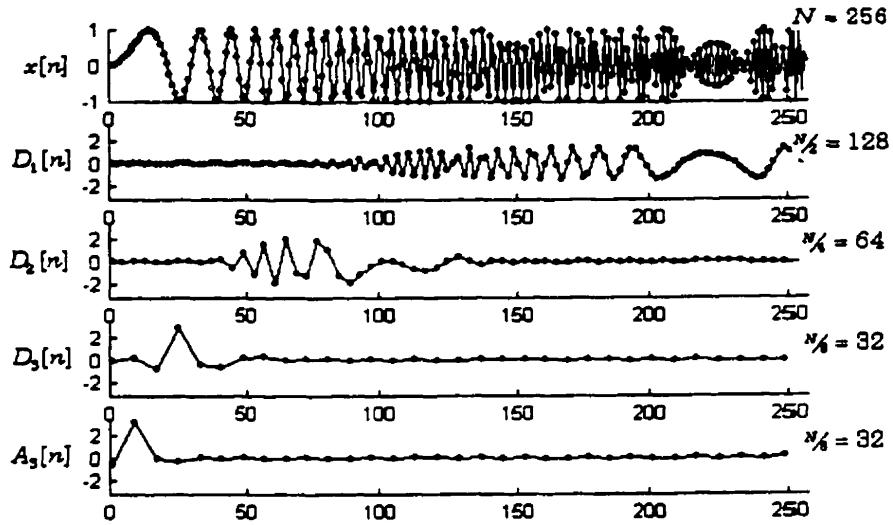


Figure 3.6 – The 3-scale decomposition of a chirp signal.

It is obvious that the high-frequency components are captured in the uppermost detail signal D_1 at the end of the time record. Progressively lower frequency components are evident in D_2 and D_3 , which occur earlier in the time record. The residual is the scale-3 lowpass approximation to the signal $x(t)$, A_3 . Note that the length of each successive level of the decomposition decreases by a factor of 2 from the original signal; the sum total of coefficients from the decomposition is always equal to N .

3.2.5 Time-Frequency Plane Tiling

One of the most important ways of interpreting the information content in a signal is through its time-frequency response. The TFR of a chirp signal using a STFT was illustrated in the previous chapter. The TFR of the WT is computed in a

slightly different manner than the STFT. More precisely, the means by which the time-frequency domain is segmented (or *tiled*) differs due to the variable tradeoff in time and frequency resolution.

Consider a signal of length $N = 2^{n_0} = 8$ ($n_0 = 3$) sampled at $f_s \text{ Hz}$. The tiling of the time-scale grid (with a wavelet decomposition to scale $J = \log_2 N = 3$) will look like that shown in Figure 3.7(a). The selection of a dyadic sampling grid ($a = 2^j$, $\tau = n2^j$) means that the resulting TFR will be *critically sampled*; there will be N discrete time-frequency cells.

For a given information cell, the temporal resolution is $\Delta t = 2^j \cdot T_s$, and the scale resolution is $\Delta a = 2^{J-j}$. The subscript 1 or 0 on the scale parameter denotes whether the subband is *detail* (bandpass) or *approximation* (lowpass). As expected, the lowest frequency subband is composed of the approximation coefficients from the highest scale level. At scale level j , there are $\frac{NT_s}{\Delta t} = N \cdot 2^{-j} = 2^{n_0-j}$ information cells. The subband at each scale $j < J$ is described by 2^{n_0-j} detail coefficients. At scale J , there is one bandpass subband described by 2^{n_0-J} detail coefficients, and one lowpass subband described by 2^{n_0-J} approximation coefficients.

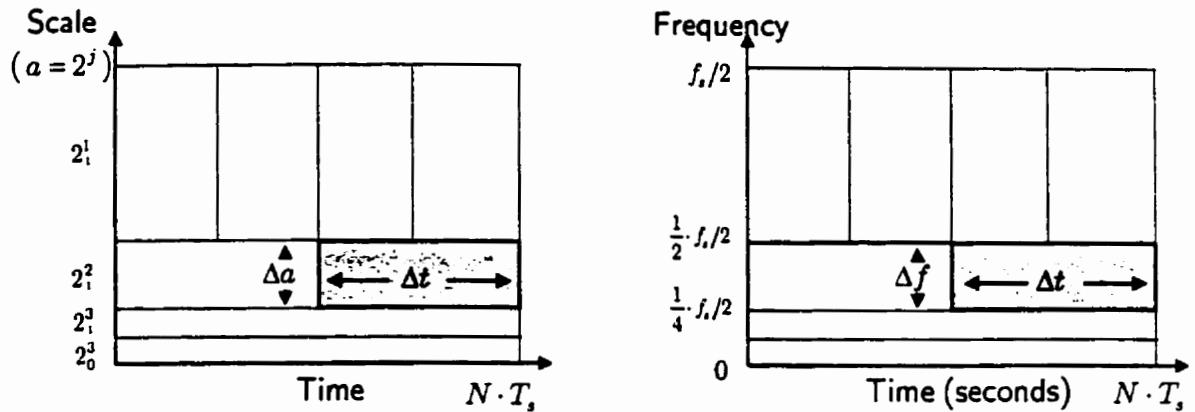


Figure 3.7 – The tiling of the time-scale domain (a) and the time-frequency domain (b) of the DWT.

Translating this time-scale response into a time-frequency response is a bit more involved. Relating now to Figure 3.7(b), the temporal resolution of a given cell is still $\Delta t = 2^j \cdot T_s$. To compute the frequency segmentation corresponding to scale, we may identify the vertical distance from 0Hz to the bottom of each cell as f , and the height of that cell as Δf . On this grid, $\Delta f = (\Delta a / 2^j)(f_s/2) = 2^{-j}(f_s/2)$. The distance f to the lower bound of a cell is given by $f = \Delta f \cdot q$, where $q = 1$ if the subband consists of *detail* coefficients and $q = 0$ if it is an *approximation* subband. Here, we have chosen a definition of Δt and Δf so as to yield a critically sampled time-frequency grid. It is interesting to note that $\Delta t \cdot \Delta f = (2^j \cdot T_s)(2^{-j} f_s/2) = 1/2$; which is somewhat greater than the minimum set by the Heisenberg bound. A chosen wavelet therefore, need not meet the Heisenberg bound but rather satisfy $\Delta t \cdot \Delta f \leq 1/2$ to avoid overlap amongst adjacent cells in the TFR⁴. This critically sampled grid also ensures perfect reconstruction of the original signal.

Consider again our chirp signal, with $N = 256$ and $f_s = 1000 \text{ Hz}$, analyzed to a depth of $J = 3$ (the maximum possible depth is $J = \log_2 256 = 8$). The two-dimensional TFR is shown in Figure 3.8(a); the corresponding three dimensional surface plot is shown in Figure 3.8(b).

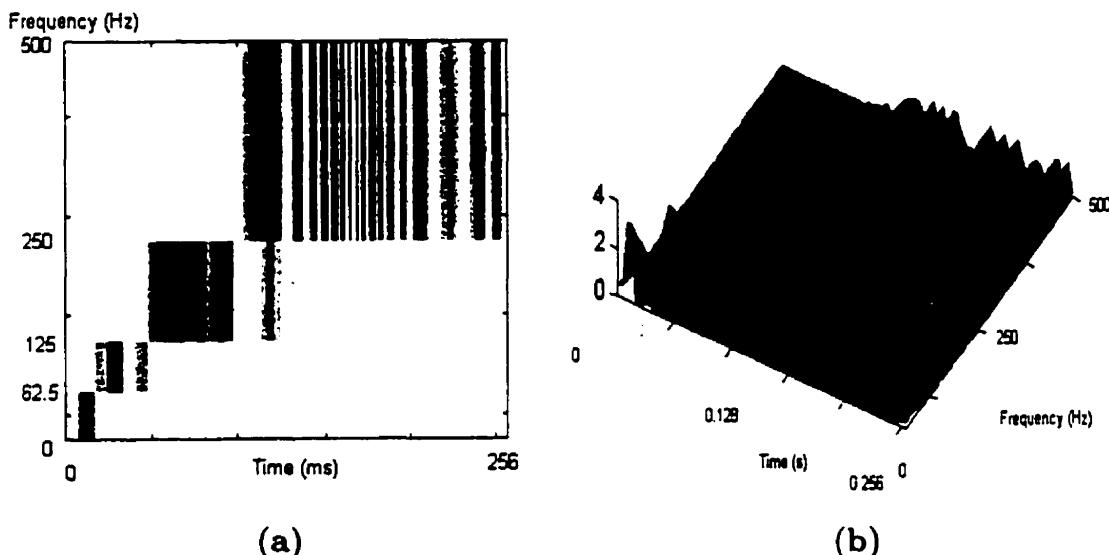


Figure 3.8 – The 3-scale DWT time-frequency response of a chirp signal.

The tiling produces finer frequency resolution toward the lower frequencies. Since the instantaneous frequency of the chirp signal increases linearly, these lower frequency (high scale) cells *localize* the response better than the high-frequency (low scale) cells, and correspondingly, have a greater amplitude.

* Most wavelets satisfy $\Delta t \cdot \Delta f \leq \frac{1}{2}$ sufficiently; any overlap is not so much as to severely distort the TFR.

3.2.6 Wavelet Selection

The selection of the basic (or *mother*) wavelet depends very much on the nature of the signals and the goal of the signal processing. Recall that the wavelet transform produces coefficients of basis vectors that form a basis in $L^2(\mathbb{R})$. The coefficients are:

$$\{D_1[n], D_2[n], \dots, D_J[n], A_J[n]\} \quad (3.24)$$

which act upon the basis vectors

$$\left\{ \{\psi_{1,n}\}_{n=0}^{2^{n_0-1}-1}, \{\psi_{2,n}\}_{n=0}^{2^{n_0-2}-1}, \{\psi_{3,n}\}_{n=0}^{2^{n_0-3}-1}, \dots, \{\psi_{J,n}\}_{n=0}^{2^{n_0-J}-1}, \{\varphi_{J,n}\}_{n=0}^{2^{n_0-J}-1} \right\} \quad (3.25)$$

each of which spans a subspace. The corresponding set of mutually orthonormal subspaces is:

$$\{\Omega_{1,1}, \Omega_{2,1}, \Omega_{3,1}, \dots, \Omega_{J,1}, \Omega_{J,0}\}. \quad (3.26)$$

From Equation (3.26), it is clear that each subspace $\Omega_{j,k}$ is spanned by 2^{n_0-j} basis vectors. To simplify matters, we may use the notation $w_{j,0,n} \doteq \psi_{j,n}$ and $w_{j,1,n} \doteq \varphi_{j,n}$, so that the set wavelet basis vectors is⁵

$$\{\{w_{1,1,n}\}, \{w_{2,1,n}\}, \{w_{3,1,n}\}, \dots, \{w_{J,1,n}\}, \{w_{J,0,n}\}\} \quad (3.27)$$

where each basis vector $w_{j,k,n} = \{w_{j,k,n}[i]\}_{i=0}^{N-1} \in \Omega_{j,k}$ for $k = 0,1$ at *scale* j and *location* n , for $0 \leq n \leq 2^{n_0-j}$, $n \in \mathbb{Z}$. Using these basis vectors, the wavelet transform may be expressed in vector-matrix form as

$$\alpha = W^T \mathbf{x} \quad (3.28)$$

⁵ This notation is offered to provide an analogy to that needed to describe wavelet packet bases, which are introduced in the next section.

where $\alpha \in \mathbb{R}^N$ contains the wavelet coefficients, and $W \in \mathbb{R}^{N \times N}$ is an orthogonal matrix consisting of the column vectors $w_{j,k,n}$. These basis vectors have the following important properties:

- (i) **vanishing moments:** $\sum_{i=0}^{L-1} i^m w_{j,k,n}[i] = 0$ for $m = 0, 1, \dots, M-1$.

The higher the degrees of vanishing moments a basis has, the better it models the smooth part of the signal. In the original wavelet basis proposed by Daubechies [Daubechies88] the length of the QMFs is given by $L = 2M$. Many other possibilities exist, for example, a family of wavelets called *Coiflets* [Daubechies92] with $L = 3M$ which are less symmetric than the original wavelets of Daubechies.

- (ii) **regularity:** $|w_{j,k,n}[i+1] - w_{j,k,n}[i]| \leq c 2^{-\rho}$,

where $c > 0$ is a constant and $\rho > 0$ is called the *regularity* of the wavelet. The larger the value of ρ , the smoother the basis vector becomes [Rioul93]. This property is important in signal compression if high ratios are desired: the shapes of the basis vectors become “visible” under these circumstances. Low regularity might result in fractal-like shapes in the reconstructed signals or images.

- (iii) **compact support:** $w_{j,k,n}[i] = 0$ for $i \notin [2^j n, 2^j n + (2^j - 1)(L - 1)]$.

This property is important for efficient and exact numerical implementation [Daubechies92].

Some wavelets are better than others for specific applications, with respect to the properties listed above. In general however, because of these properties, wavelet bases generate very efficient and simple representations for piecewise smooth signals and images. The manner in which vanishing moments, regularity and

compact support affect the wavelet's efficacy as a basis for signal classification is not clear. One would expect that a wavelet that "looks like" the elemental components of the signals under consideration would be the most appropriate. Most commonly used wavelets *do* resemble the elemental structure in the MES – the motor unit action potential. For a given wavelet, it is reasonable to expect that the small scales would capture isolated motor unit activity, while larger scales would model longer-duration trends in the signal. More important however, is the ability of the wavelet basis to generate a TFR that clearly distinguishes signals in different classes. This requires that the wavelet functions appropriately model the signal, and that they be well localized and well behaved in the time-frequency plane.

3.3 Wavelet Packet and Cosine Packet Bases

For many signals of physical origin (especially those that are discontinuous or transient), the WT possesses time-frequency localization that is superior to the STFT due to its dyadic tiling. Its time-frequency tiling is, nonetheless, still fixed. The wavelet packet transform (WPT) and cosine packet transform (CPT) permit an *adaptive* time-frequency tiling. These decompositions produce an overcomplete set of subspaces, making it possible to select one of many coordinate systems with which the signals may be “viewed”. Using a cost function designed for a specific signal processing goal (for example, compression or classification) the *best basis* can be chosen to *optimize* the coordinate system with respect to frequency (*via* wavelet packet bases) or time (*via* cosine packet bases).

3.3.1 The Wavelet Packet Transform

The *wavelet packet transform* [Coifman89][[Coifman90][Meyer93][Wickerhauser94] is a generalized version of the wavelet transform: it retains not only the low but also the high frequency subband, performing a decomposition upon both at each stage. As a result, the tiling of the time-frequency plane is configurable: the partitioning of the frequency axis may take many forms to suit the needs of the application. This is demonstrated in Figure 3.9, which has been reproduced from Chapter 2, for convenience.

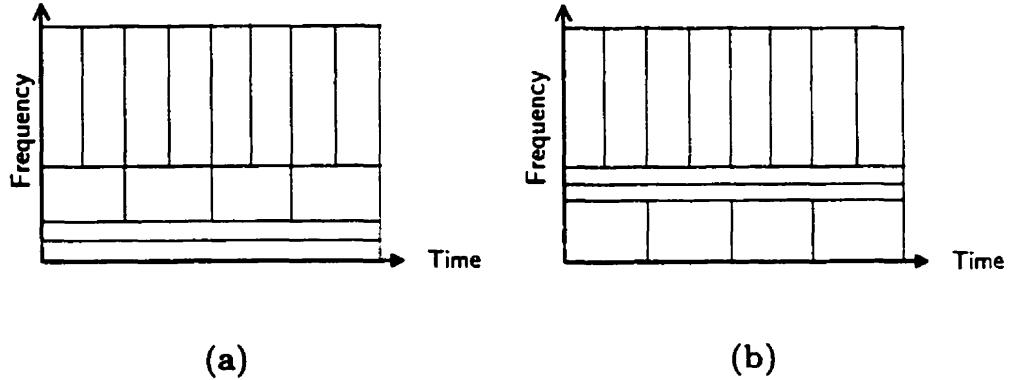


Figure 3.9 – The time-frequency plane tiling of
 (a) a wavelet basis, and (b) an arbitrary wavelet packet basis.

Starting with a signal \mathbf{x} of length N samples, the first level of the decomposition generates the lowpass and highpass subbands ($H\mathbf{x}$ and $G\mathbf{x}$ respectively) as with the wavelet transform, each of half the length of \mathbf{x} . The second level decomposition generates four subsequences: $H^2\mathbf{x}, GH\mathbf{x}, HG\mathbf{x}, G^2\mathbf{x}$ halved in length again; the decomposition to this level is shown in Figure 3.10.

| \mathbf{x} | | | |
|-----------------|----------------|----------------|-----------------|
| $H\mathbf{x}$ | | $G\mathbf{x}$ | |
| $H^2\mathbf{x}$ | $GH\mathbf{x}$ | $HG\mathbf{x}$ | $G^2\mathbf{x}$ |

Figure 3.10 – The first 2 levels of decomposition in a wavelet packet transform.
 The lowpass operations (H) occur to the left, the highpass (G) to the right.

This process is repeated J times, where $J \leq \log_2 N$, resulting in JN coefficients. The computational cost of this decomposition is on the order of $O(JN) \leq O(N \log_2 N)$ [Wickerhauser94]. This iterative process generates a *binary wavelet packet tree* structure where the nodes of the tree represent subspaces with different frequency localization characteristics. The binary wavelet packet tree for a Kronecker delta function is shown in Figure 3.11. Each subband in the full

decomposition is separated by dashed lines; the wavelet packet coefficients within each subband are shown as solid lines. The vertical axis indicates the depth of decomposition; the zero scale is actually the original signal. The horizontal axis is labeled **Frequency[Time]** implying that, for a given scale (level), the centre frequency localization of each subband increases as one progresses from left to right. Within each subband, the horizontal axis reflects time: for a subband at scale j , there are $N/2^j$ wavelet packet coefficients sampled at $f_s/2^j$ spanning the entire record NT .

Now consider the characteristics of the Kronecker delta function, which is zero at all locations except a single sample (as indicated at level 0). At the singularity, the signal (theoretically) contains all frequencies. Each subband in the decomposition contains a nonzero coefficient at the temporal location of the singularity⁶.

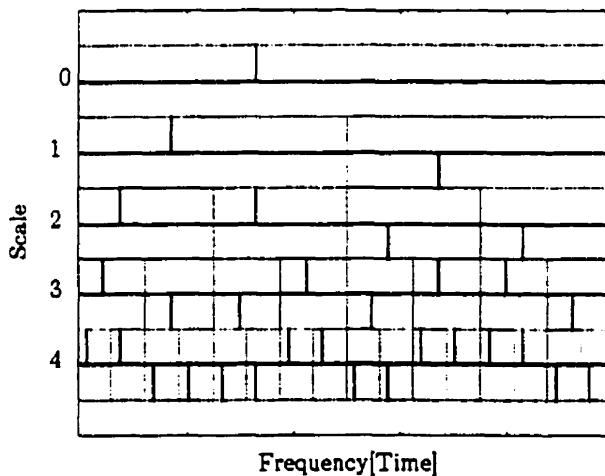


Figure 3.11 – The binary wavelet packet tree for a Kronecker delta function, using a Haar wavelet packet basis. The vertical axis represents the depth (level) of the wavelet packet decomposition, the horizontal axis corresponds to frequency localization between subbands, and to time localization within a given subband.

The Kronecker delta is the epitome of a temporally localized signal. Consider now the case of a signal ideally localized in frequency: a single sinusoid. The binary

⁶ This decomposition was performed using Haar wavelets (rectangular pulses) which are ideally localized in time. This eliminates any "sideband" coefficients that would exist if the wavelet could not characterize the singularity with a single basis function in each subband.

wavelet packet tree is shown in Figure 3.12. It is evident that the localization of the signal becomes more resolved as one progresses down the tree, and is represented as a single tone at $J=5$.

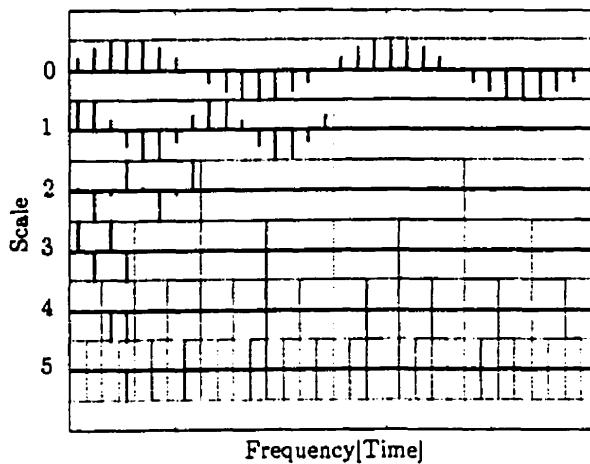


Figure 3.12 – The binary wavelet packet tree for a sinusoid of $N=32$ samples. A Daubechies-20 wavelet packet basis was used, which provides good frequency localization.

A similar analysis upon the linear chirp signal yields the binary wavelet packet tree depicted in Figure 3.13. The time-frequency localization of each subband is evident; within each subband, the significant coefficients are temporally localized in the range where the chirp passes through the frequency range corresponding to that subband.

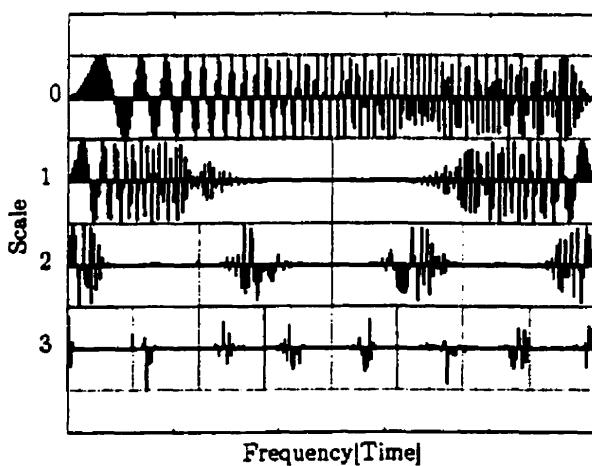


Figure 3.13 – The binary wavelet packet tree for a 256-sample chirp signal. A Coiflet-5 wavelet was used, which provides a good tradeoff between time and frequency localization.

These examples are simple, but convey the notion of the wavelet packet decomposition, and the overcomplete representation in the form of a binary tree.

In subspace notation, the root node of the tree is $\Omega_{0,0}$. The node $\Omega_{j,k}$ is decomposed into two orthogonal subspaces $H : \Omega_{j,k} \rightarrow \Omega_{j+1,2k}$ and $G : \Omega_{j,k} \rightarrow \Omega_{j+1,2k+1}$. Here j denotes scale, as before, and k indicates the subband index within the scale⁷. This may be expressed as

$$\Omega_{j,k} = \Omega_{j+1,2k} \oplus \Omega_{j+1,2k+1} \text{ for } j = 0, \dots, J \text{ and } k = 0, \dots, 2^j - 1 \quad (3.29)$$

which generates the entire decomposition to scale J . A decomposition to scale $J = 3$ is shown in Figure 3.14.

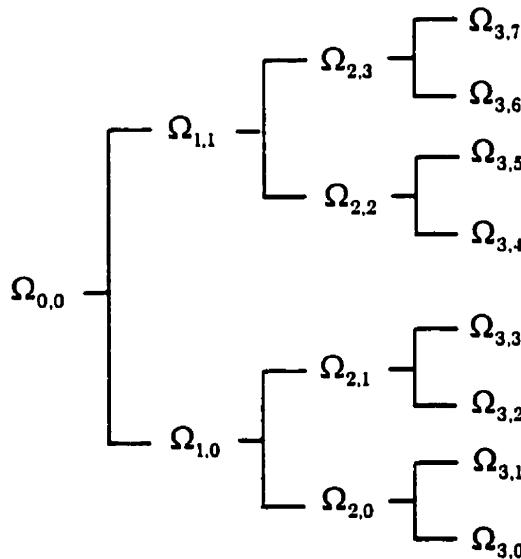


Figure 3.14 – A decomposition of $\Omega_{0,0}$ into binary tree-structured subspaces using the wavelet packet transform (with $J=3$).

The set of subspaces in the binary tree is a redundant set. Indeed, there are more than $2^{2^{J-1}}$ possible orthonormal bases in this binary tree [Wickerhauser94]. Each

⁷ The wavelet transform has only two subbands per scale, high and low, with $k=0,1$.

subspace $\Omega_{j,k}$ is spanned by 2^{n_0-j} basis vectors $\{w_{j,k,n}\}_{n=0}^{2^{n_0-j}-1}$. In a given family of wavelet packet bases, the parameter j indicates scale, as before. The parameters k and n roughly indicate frequency band⁸ and the centre of the waveform, respectively. The vector $w_{j,k,n}$ is roughly centred at $2^j n$, has length of support $\approx 2^j$, and oscillates $\approx k$ times. For $j=0$, we have the original signal space (the standard Euclidean basis \mathbb{R}^N). A set of typical wavelet packet basis vectors is shown in Figure 3.15.

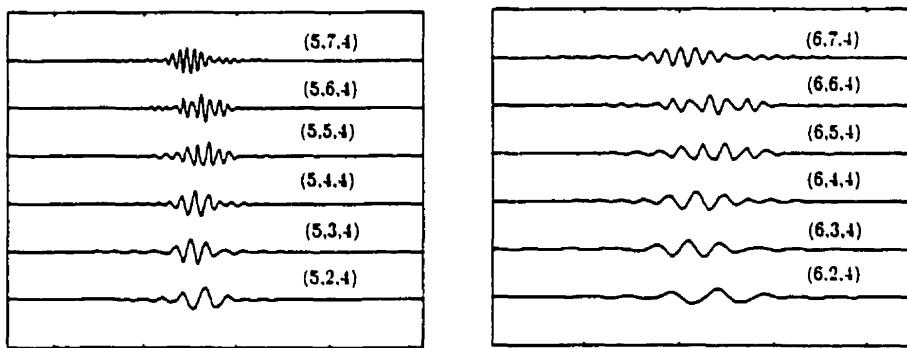


Figure 3.15 – Some wavelet packet basis vectors for scale=5 and scale=6.
All packets are indexed as $(j, k, n) = \text{scale, frequency, location}$. For the same location and scale, the oscillation frequency increases with k .

The question now is: how does one find the *best* basis for the problem at hand, from this overcomplete set of bases? Best basis selection for signal compression and signal classification is the subject of Section 3.4.

⁸ The binary tree subband structure, as generated by successive applications of H and G is called *Paley* or *natural* ordering. In this form, the frequency band of $\Omega_{j,k}$ does not monotonically increase with k . This may be corrected by *Gray-code* permutation [Saito94].

3.3.2 The Cosine Packet Transform

Local trigonometric transforms [Coifman91][Auscher92] can be regarded as the conjugate operation of a WPT. Whereas wavelet packets partition the frequency axis, local trigonometric transforms partition the time axis. Indeed, while the WPT yields an orthonormal basis on each subband, the local trigonometric transform yields an orthonormal basis on each temporal interval. This is demonstrated in Figure 3.16, which shows an example of tiling in the time-frequency plane. The most common forms are the *local cosine packet transform* (CPT) and the *local sine packet transform*, which are explained in detail later in this section.

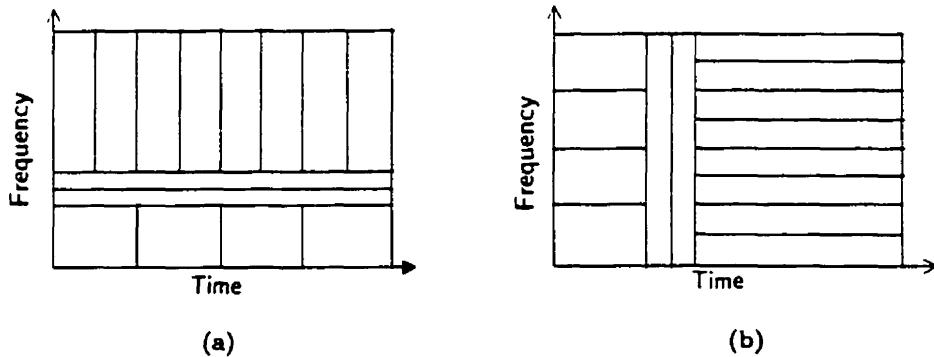


Figure 3.16 – The time-frequency plane tiling of
 (a) a wavelet packet basis, and b) a local trigonometric basis.

The basis functions actually overlap between intervals⁹, thus the time axis is partitioned smoothly: the local Fourier analyses on each interval have fewer edge effects than discrete cosine or sine transforms. Local trigonometric transforms produce the same binary packet tree as wavelet packet transforms, except that the *subbands* are instead *intervals* because each subspace in the decomposition is partitioned with respect to *time* instead of *frequency*. The binary cosine packet tree for a Kronecker delta function is shown in Figure 3.17.

⁹ Hence, they are sometimes referred to as *lapped orthogonal transforms* [Malvar90].

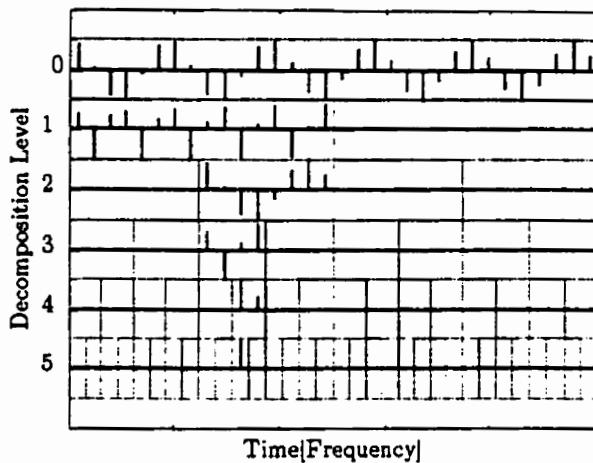


Figure 3.17 – The binary cosine packet tree for a Kronecker delta function ($N=32$). The vertical axis represents the depth (level) of the cosine packet decomposition, the horizontal axis corresponds to time localization with respect to intervals, and to frequency localization within a given interval. For level $j=0$, the CP coefficients are the DCT-IV coefficients.

Again, the vertical axis represents the depth of decomposition, but now, instead of partitioning the frequency axis, the decomposition partitions the time axis. The horizontal is labeled Time[Frequency]; what this implies is that intervals are indexed with respect to time, and the coefficients within each interval are ordered with respect to frequency¹⁰.

Clearly, as the decomposition (temporal partitioning) of the Kronecker delta signal progresses, only a single interval contains significant energy; at each level the isolation of the singularity is more resolved in time.

Consider now the cosine packet basis functions themselves: some typical cosine packet basis functions are shown in Figure 3.18. The basis function indexing (j, k, n) now must be regarded as scale, location (or window index) and frequency, respectively.

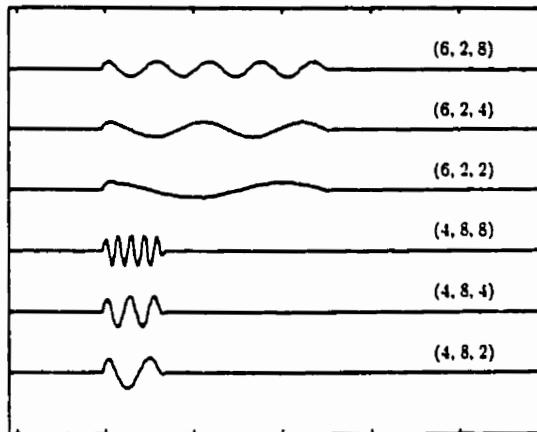


Figure 3.18 – Some cosine packet basis functions. The index (j,k,n) implies scale, location and frequency, respectively.

It is evident that scale, frequency, and localization have the same manifestation as they demonstrated for wavelet packets. The cosine/sine packet bases are essentially localized (or *windowed*) sinusoids. To achieve temporal localization, a particular symmetric window (or *bell*) function is used:

$$b(t) \doteq \begin{cases} \sin \frac{\pi}{4}(1 + \sin \pi t) & \text{if } -\frac{1}{2} < t < \frac{3}{2}, \\ 0 & \text{otherwise.} \end{cases} \quad (3.30)$$

This is shown in Figure 3.19.

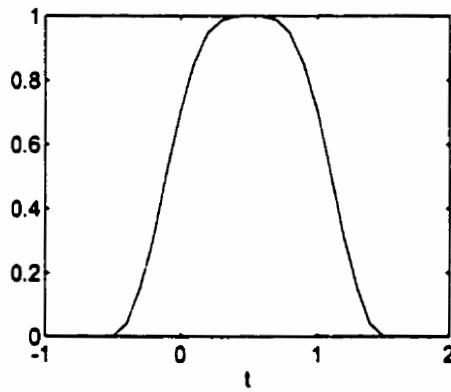


Figure 3.19 – A bell function used to achieve temporal localization.

¹⁰ The response at level zero is the discrete cosine transform (DCT-IV) upon the original signal; in each interval at successive levels of decomposition, the response is a modified (by scaling and “folding” the basis functions) form of the DCT-IV [Auscher92].

Consider the interval of integers $I = \{0, 1, \dots, N-1\} \equiv [0, N)$. The local trigonometric functions are mainly supported on I but also take on some values on $\{-N/2, \dots, -1\}$ and $\{N, \dots, 3N/2\}$. For integers $k \in I$ the *local sine basis functions* are given by

$$S_k(n) \doteq \frac{1}{\sqrt{2N}} b \left(\frac{n + \frac{1}{2}}{N} \right) \sin \left(\pi \left(k + \frac{1}{2} \right) \left(\frac{n + \frac{1}{2}}{N} \right) \right). \quad (3.31)$$

Here, k becomes a location (or window index) parameter, and n is a measure of frequency. By replacing $\sin(\)$ by $\cos(\)$, we obtain the *local cosine basis function* $C_k(n)$. The bell function allows sines on adjacent intervals to overlap while remaining orthogonal [Auscher92].

A record may be temporally segmented into arbitrary disjoint intervals, so that the interval may be segmented into dyadic intervals, recursively. If the original record has $N = 2^{n_0}$ samples, each subinterval has 2^{n_0-j} samples at step j . Thus, the original interval $I = [0, N)$ is split into $[0, N/2)$ and $[N/2, N)$; each successive level of the decomposition splits the subintervals into half. Let $I_{0,0} = I$ be the index of the original record, and let $I_{j,k}$ be a subinterval of I after j iterations of the decomposition. The following relationship results:

$$I_{j,k} = I_{j+1,2k} \cup I_{j+1,2k+1} \quad (3.32)$$

for $j = 0, 1, \dots, J$, and $k = 0, 1, \dots, 2^j - 1$. Each interval $I_{j,k}$ may be associated with a subspace $\Omega_{j,k}$, meaning that a binary tree of the subspaces can be compiled, having the same tree structure as demonstrated for wavelet packets. Each temporal subspace $\Omega_{j,k}$ is spanned by the set of basis vectors $\{w_{j,k,n}\}_{n=0}^{2^{n_0-j}-1}$ which cover all frequencies (indexed by n), where

$$w_{j,k,n}(m) \doteq \frac{1}{\sqrt{2^{n_0-j+1}}} b \left(\frac{m + \frac{1}{2}}{2^{n_0-j}} - k \right) \sin \left(\pi \left[n + \frac{1}{2} \left\lceil \frac{m + \frac{1}{2}}{2^{n_0-j}} - k \right\rceil \right] \right). \quad (3.33)$$

for the sine packet basis. The cosine packet basis has the same form with $\sin()$ replaced by $\cos()$. The difference between a sine and cosine packet basis is a matter of triviality; the term *cosine packet transform* will be used henceforth to mean local trigonometric transform, unless otherwise specified. The computational complexity associated with this decomposition is $O(N(\log_2 N)^2)$ [Wickerhauser94]. The question, again, is how to obtain the best basis from the redundant set of bases in the binary tree. This is addressed in the next section.

3.4 Best Basis Selection

The last two sections have introduced two complementary means of decomposing a signal into an hierarchy of dyadic subspaces: the frequency-oriented wavelet packet transform, and the temporally oriented cosine packet transform. For a J -scale decomposition, the resulting binary tree yields more than $2^{2(J-1)}$ orthonormal bases (or coordinate systems), all of which offer a complete description of the space of the original signal. The power of wavelet packet and cosine packet transforms is that a “best basis” can be chosen for a specific task, if it can be properly identified from the ensemble of possible candidates.

To determine the *best basis*, it is necessary to evaluate and compare the efficacy of many bases. To this end, a *cost function* must be chosen to represent the goal of the application. The best-basis selection algorithm has its origins in signal compression [Coifman92][Wickerhauser94], and the cost functions associated with compression all entail some form of entropy measure. This form of the best basis algorithm is the simplest, and will be used to introduce the concept of best basis selection. Subsequently, it will be shown how the algorithm may be modified to suit the classification problem. The best basis algorithm operates on the binary tree of subspaces; it does not need to know whether the subspaces are oriented in frequency or in time, so that it works identically with both wavelet packet and cosine packet transforms.

3.4.1 Best Basis Selection for Signal Compression

The best-basis selection algorithm operates on a *single* signal, or more specifically, its binary packet tree of orthonormal bases. The *best basis algorithm* proposed by Coifman and Wickerhauser [Coifman92] [Wickerhauser94] is a divide-and conquer search of the binary tree (a *pruning algorithm*) in which one begins with a fully decomposed tree, starts at the lowest level, and eliminates branches until an optimal solution is found¹¹.

The cost function associated with the pruning algorithm is based on entropy since, for signal compression, the goal is to maximize the information with respect to the chosen set of coordinate axes. A natural choice is the *Shannon entropy*:

$$H(\mathbf{p}) = \sum_i p_i \log_2 p_i, \quad (3.34)$$

where $\mathbf{p} = \{p_i\}$ is a nonnegative sequence with $\sum_i p_i = 1$. Certainly, other entropy-based measures are possible [Wickerhauser94] with varying effects on the outcome of the algorithm. These will not be considered here.

What follows is a brief description of the pruning algorithm. Consider a single subspace $\Omega_{j,k}$ within a binary packet tree. Let $B_{j,k}$ denote a set of basis vectors belonging to the subspace $\Omega_{j,k}$, arranged in matrix form:

$$B_{j,k} = [\mathbf{w}_{j,k,0}, \mathbf{w}_{j,k,1}, \dots, \mathbf{w}_{j,k,2^{n_j-1}}]^T. \quad (3.35)$$

Let $A_{j,k}$ represent the best basis for the signal \mathbf{x} restricted to the span of $B_{j,k}$, and let \aleph be the chosen information cost function. The following algorithm “prunes”

¹¹ An alternative has been suggested [Taswell95] in the form of a *growing algorithm* which expands the tree as needed, but this does not always reach an optimal solution since the entire tree may not be searched.

the binary tree by comparing the cost function of each parent node with its two children.

Algorithm 3.1 (The Best Basis Algorithm)[Wickerhauser94].

Given a signal \mathbf{x} ,

Step 0: Choose a time-frequency decomposition method. That is, specify a wavelet packet transform (i.e. a pair of QMFs) or a cosine packet transform. Specify the depth of decomposition J , and an information cost function \aleph .

Step 1: Decompose \mathbf{x} into its binary packet tree, and obtain the coefficients $\{B_{j,k}\}$ for $0 \leq j \leq J$ and $0 \leq k \leq 2^j - 1$.

Step 2: Begin at level J : set $A_{J,k} = B_{J,k}$ for $k = 0, \dots, 2^J - 1$.

Step 3: Determine the best subspace $A_{j,k}$ for $j = J-1, \dots, 0$, $k = 0, \dots, 2^j - 1$ by

$$A_{j,k} = \begin{cases} B_{j,k} & \text{if } \aleph(B_{j,k}\mathbf{x}) \leq \aleph(A_{j+1,2k}\mathbf{x} \cup A_{j+1,2k+1}\mathbf{x}) \\ A_{j+1,2k} \oplus A_{j+1,2k+1} & \text{otherwise.} \end{cases} \quad (3.36)$$

When the algorithm has completed, we are left with $A_{0,0}$, which is the best basis for the signal \mathbf{x} restricted to the span of $B_{0,0} \equiv \mathfrak{R}^N$. The chosen best basis consists of a disjoint set of subspaces, and each subspace $\Omega_{j,k}$ contains 2^{n_0-j} basis vectors. The total number of basis functions is always N , where $N = 2^{n_0}$ is the length of each signal \mathbf{x} . To make this algorithm fast, the cost function \aleph must be *additive*: $\aleph(\{\mathbf{x}_i\}) = \sum_i \aleph(\mathbf{x}_i)$ so that

$$\aleph(A_{j+1,2k}\mathbf{x} \cup A_{j+1,2k+1}\mathbf{x}) = \aleph(A_{j+1,2k}\mathbf{x}) + \aleph(A_{j+1,2k+1}\mathbf{x}). \quad (3.37)$$

This implies that a simple addition suffices instead of computing the cost of the union of the nodes. The proof that this algorithm yields the best basis relative to

an additive form of \mathbf{N} may be found in [Coifman92]. The computational complexity of the best basis algorithm is $O(N)$. Given this best basis, the transform must find only the coefficients corresponding to the chosen subspaces, instead of all coefficients in the overcomplete binary packet tree.

Once a best basis has been chosen, it is desirable to have a means to interpret its meaning with respect to time-frequency localization. This is best illustrated by means of an example; for this we once again turn to the linear chirp signal. Consider first a wavelet packet analysis. Using a Coiflet-3 wavelet packet QMF [Coifman92], the wavelet packet binary tree was computed, shown at the top left of Figure 3.20. Using a Shannon entropy cost metric, the best basis was found using Coifman and Wickerhauser's pruning algorithm. One possible way of visualizing this best basis is by its *basis tree*, as depicted in the top right of Figure 3.20. The branches of the tree depict the segmentation of the frequency axis using the chosen basis; the height of each branch is proportional to the improvement in the cost function obtained by splitting the parent node into its children.

The selection of the best basis also specifies the tiling of the time-frequency plane, since the best basis is a set of mutually orthogonal bases. The time-frequency tiling for the chirp signal is shown in the bottom left of Figure 3.20. Directly related to the time-frequency tiling is the actual time-frequency response. The image plot of the chirp signal's time-frequency response is also shown in the bottom right of Figure 3.20. The mathematics of computing the time-frequency cells in the wavelet packet and cosine packet transform are a bit more involved than with the wavelet transform. The details have been included in Appendix C.

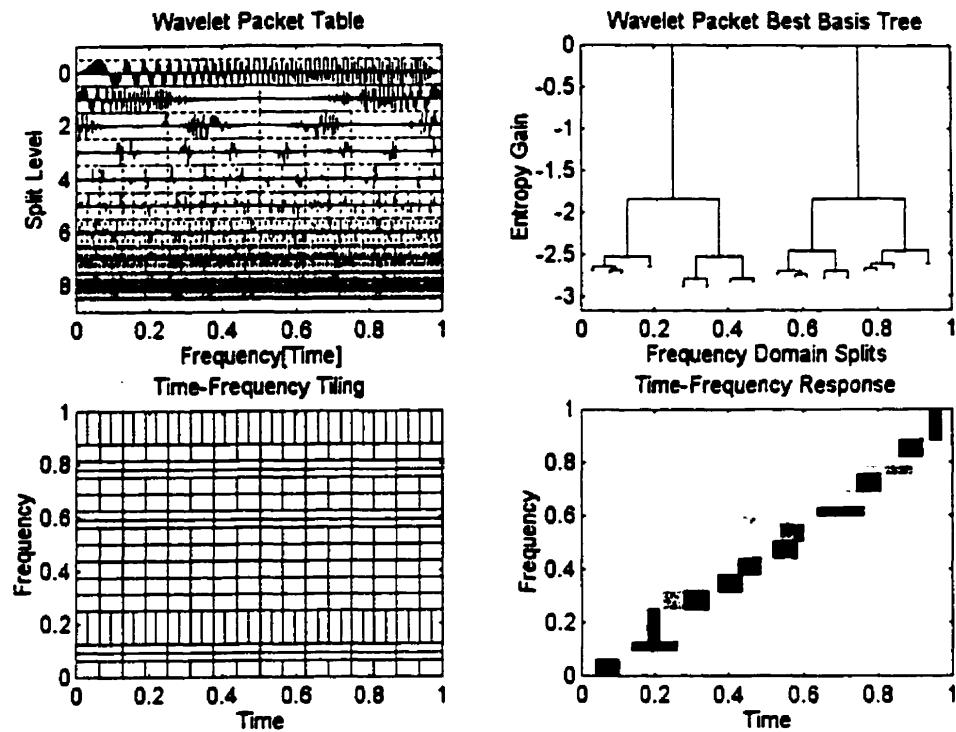


Figure 3.20 – The characteristics of the wavelet packet best basis chosen for a chirp signal, subject to entropy-based compression.

A cosine packet analysis was performed upon the same chirp signal, using the same Shannon entropy cost function. The cosine packet binary table, the best-basis tree, the time-frequency tiling, and the time-frequency response are shown in Figure 3.21. It is evident that wavelet packets and cosine packets partition the time-frequency plane in a complementary sense.

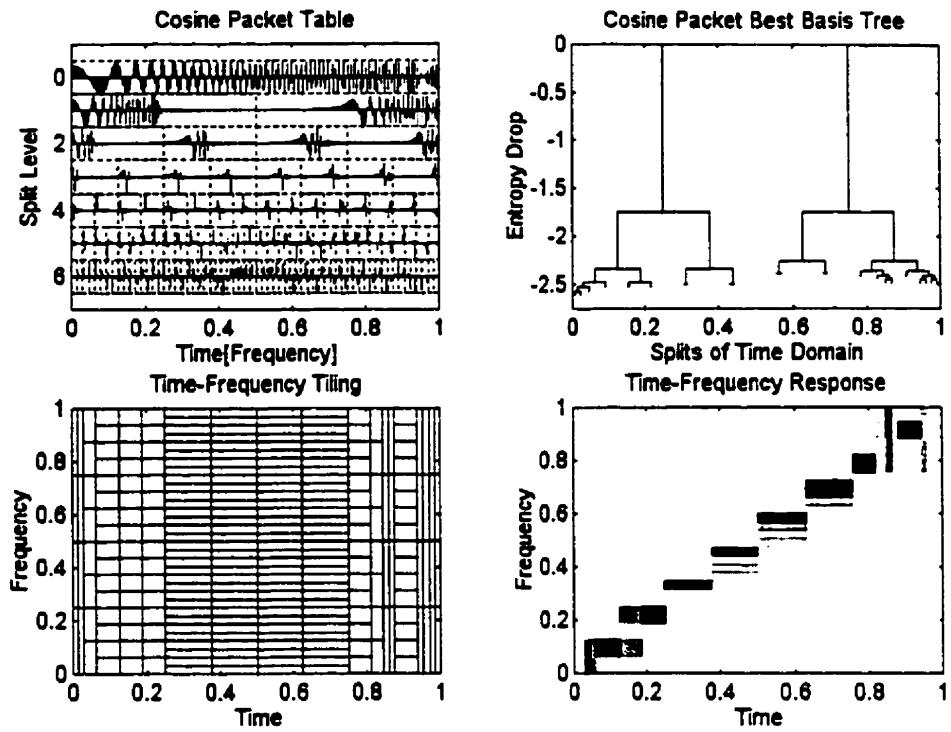


Figure 3.21 – The characteristics of the cosine packet best basis chosen for a chirp signal, subject to entropy-based compression.

For signal compression, the next step after computing the coefficients of the best basis would be to threshold the coefficients; retaining only as many needed to reconstruct the signal to some specified measure of quality. Finally, the original signal is reconstructed from this subset of the transform coefficients. In a hardlimiting thresholding scheme, only the coefficients which exceed a threshold are kept. The “goodness of fit” can be measured by the standard error:

$$\epsilon = (\sum (x - \hat{x})^2) / (\sum x^2),$$

where x is original signal, and \hat{x} the synthesis, reconstructed from the thresholded coefficients. Figure 3.22 shows the standard error resulting from the analysis/threshold/synthesis of the 256-point linear chirp, using the WT, WPT, and CPT. The standard error is plotted against the number of coefficients used in the reconstruction. The compression ratio is the ratio of the original number of points ($N=256$) to the number of coefficients. Clearly, the

wavelet packet representation is superior to the others, and the wavelet method is inferior to both of the adaptive (packet-based) techniques.

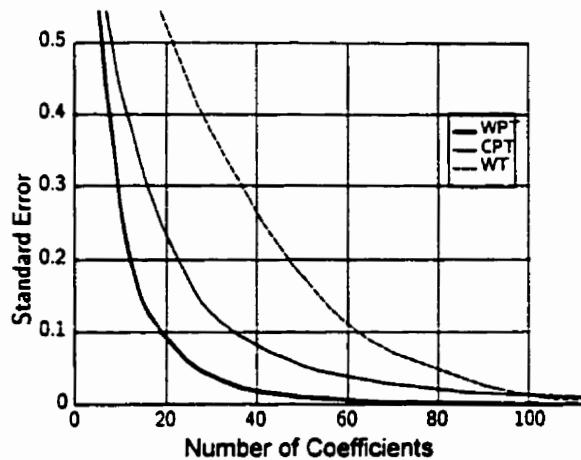


Figure 3.22 – The standard error in reconstructing a linear chirp signal using the WT, WPT, and CPT.

Figure 3.23 depicts the standard error in reconstructing a 256-sample record corresponding to elbow flexion, using the WT, WPT, and CPT. Again, the packet transforms outperform the WT, and the WPT provides the most efficient basis for reconstructing this pattern.

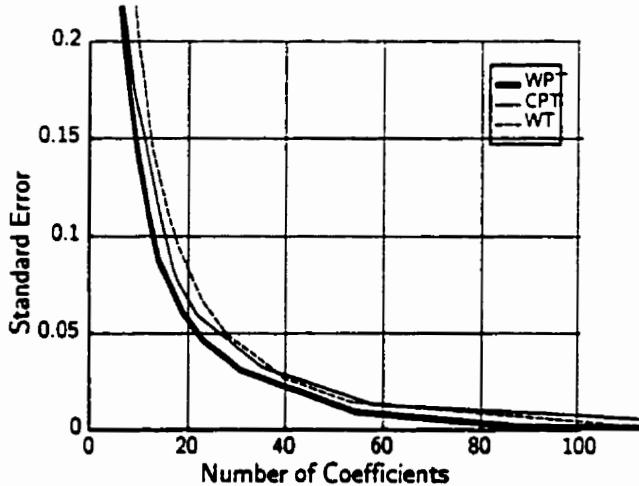


Figure 3.23 – The compression of a transient MES pattern ($N=256$) corresponding to elbow flexion. The figure depicts the standard error corresponding to wavelet, wavelet packet, and cosine packet transforms.

Consider a point on the standard error curve corresponding to WPT compression, such that 16 (of 256) WPT coefficients are used in the reconstruction. This yields

a 16:1 compression ratio with a standard error of 0.134. The elbow flexion signal and the reconstructed waveform are shown in Figure 3.24.

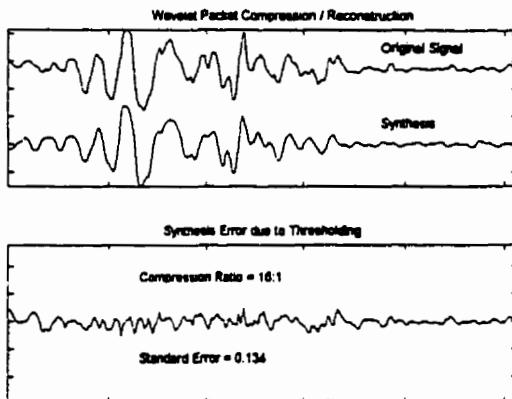


Figure 3.24 – The original signal, the reconstructed signal and the synthesis error of an elbow flexion pattern subject to WPT compression.

It is evident that regeneration of the fundamental shape of the waveform is very good. It seems that the fine “noise” detail in the waveform is lost during compression/synthesis. This is but a simple example of wavelet-based compression, but it demonstrates the power of wavelet based representation in modeling the transient MES. It is intuitively pleasing to find that the transient MES is well-modeled by wavelet bases since the wavelets themselves resemble the elemental structures of the MES (the MUAPs). In this sense, the transform provides a parametric model of the transient patterns, in the sense that it determines the set of shifted and scaled basis functions that yields the best fit to the volley of MUAPs.

3.4.2 Best Basis Selection for Classification

Fundamental to the success of any classifier is the quality of the feature set with which it is provided. The desirable properties of a feature set for classification are:

1. the *statistical distances* between classes are maximized, and
2. the feature set supplies the most important features, and suppresses the redundant ones.

These concepts were emphasized in Chapter 2. Furthermore, it was proposed that the time-frequency domain provides a revealing description of transient signals. The promise of TFRs as a basis for feature extraction lies in their ability to concentrate (or localize) information that would otherwise be dispersed in either time or frequency alone. Wavelet and cosine packet transforms offer an advantage over the STFT and the WT in that they provide an adaptive time-frequency tiling via best basis selection. The previous section demonstrated how an adaptive basis results in signal compression properties that are superior to fixed bases. This section presents an algorithm for selecting the “best basis” for signal classification.

3.4.2.1 Discriminant Measures

In order to determine the best basis for classification amongst the ensemble of redundant bases in a complete packet decomposition, it is necessary to establish a measure of discriminant power. As explained in Chapter 2, the ideal criterion would be the probability of misclassification, evaluated upon each candidate basis. In practice, evaluation of this criterion is generally too complex, and one must resort to simpler criteria, such as class separability. Additionally, an ideal evaluation would have each of the 2^N possible orthonormal bases compared in terms of discriminant power. A suboptimal technique that requires far less

computation is to *prune* the packet tree by evaluating the individual discriminability of each subband. This pruning algorithm is discussed in the next section.

For now, it will be assumed that class separability is the most practical measure of discriminant power. If we are to evaluate the discriminability of candidate subbands upon these terms, we are faced with the same problem as *feature selection for dimensionality reduction*.

Recall from Section 2.6.1 that an n -feature discriminant measure can be defined as $D(\mathbf{p}, \mathbf{q})$, where $\mathbf{p} = \{p_i\}_{i=1}^n$, $\mathbf{q} = \{q_i\}_{i=1}^n$ are measures used to represent the n features. If p_i and q_i are scalars (e.g. the mean energy of the i^{th} feature) then the discriminant measure may take one of the following forms¹²:

(i) **Relative Entropy:**

$$D(\mathbf{p}, \mathbf{q}) \doteq \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \quad (3.38)$$

which measures the discrepancy of \mathbf{p} from \mathbf{q} [Kullback51]. The drawback to this measure is that it is not symmetric in \mathbf{p} and \mathbf{q} : characteristics of the features in \mathbf{p} with respect to \mathbf{q} will not yield the same measure if the class order is reversed. This may tend to *bias* the relative entropy measure to the activity in one class over another. This is desirable if the goal is to separate signal from noise, but does not give a fair treatment amongst classes in pattern recognition.

¹² Each of the following discriminant measures exploit the fact that they are additive.

(ii) **Symmetric Relative Entropy:**

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) \doteq \sum_{i=1}^n p_i \log \frac{p_i}{q_i} + \sum_{i=1}^n q_i \log \frac{q_i}{p_i} \quad (3.39)$$

which does yield symmetric activity amongst classes: clearly, $\mathcal{D}(\mathbf{p}, \mathbf{q}) = \mathcal{D}(\mathbf{q}, \mathbf{p})$.

(iii) **Euclidean Distance:**

$$\mathcal{D}(\mathbf{p}, \mathbf{q}) \doteq \|\mathbf{p} - \mathbf{q}\| = \sum_{i=1}^n (p_i - q_i)^2 \quad (3.40)$$

which is another asymmetric measure [Wantabe85].

Obviously, it is necessary to discriminate amongst more than two classes. To compute the discrepancy between the distributions of K classes: $\mathbf{p}^{(1)}, \mathbf{p}^{(2)}, \dots, \mathbf{p}^{(K)}$.

one must take $\binom{K}{2}$ pairwise combinations of \mathcal{D} [Saito95]:

$$\mathcal{D}\left(\{\mathbf{p}^{(k)}\}_{k=1}^K\right) \doteq \sum_{i=1}^{K-1} \sum_{j=i+1}^K \mathcal{D}(\mathbf{p}^{(i)}, \mathbf{p}^{(j)}) \quad (3.41)$$

3.4.2.2 The Local Discriminant Basis Algorithm

The best basis algorithm, originally developed for compression, was modified to suit classification by Naoki Saito in his Ph.D. dissertation [Saito94], under the supervision of Dr. Ronald Coifman at Yale University. He termed the algorithm the *local discriminant basis* (LDB) algorithm, implying that an orthonormal basis is selected from the binary wavelet (or cosine) packet tree which most discriminates data from a given set of classes.

The measure of class separability is conveyed by the discriminant measure \mathcal{D} . In order to optimize the representation with respect to the time-frequency localization characteristics of the wavelet (or cosine) packet basis, the input parameters to \mathcal{D} are the *time-frequency energy maps* of each class.

Definition: Let $\{\mathbf{x}_i^{(c)}\}_{i=1}^{N_c}$ be a set of training signals belonging to class c , where N_c is the number of patterns in class c . The **time-frequency energy map** of class c is a table of positive real values indexed by (j, k, n) :

$$\Gamma_c(j, k, n) \doteq \frac{\sum_{i=1}^{N_c} (\mathbf{w}_{j,k,n}^T \mathbf{x}_i^{(c)})^2}{\sum_{i=1}^{N_c} \|\mathbf{x}_i^{(c)}\|^2}, \quad (3.42)$$

for $j = 0, \dots, J$, $k = 0, \dots, 2^j - 1$, $n = 0, \dots, 2^{n_0-j} - 1$. That is, Γ_c is computed by accumulating the squares of the transform coefficients for each entry in the binary packet tree (j, k, n) , and normalizing by the total energy of the signal belonging to class c .

Since the algorithm must choose the best set of subspaces from the binary packet tree, the response from individual temporal locations from within a subspace must be summed. For K classes, the overall discriminant measure for the subspace $\Omega_{j,k}$ is thus:

$$\mathcal{D}\left(\{\Gamma_c(j, k, \bullet)\}_{c=1}^K\right) \doteq \sum_{n=0}^{2^{n_0-j}-1} \mathcal{D}(\Gamma_1(j, k, n), \dots, \Gamma_K(j, k, n)). \quad (3.43)$$

The packet decomposition into a series of energy maps suitable for this discriminant measure is illustrated in Figure 3.25.

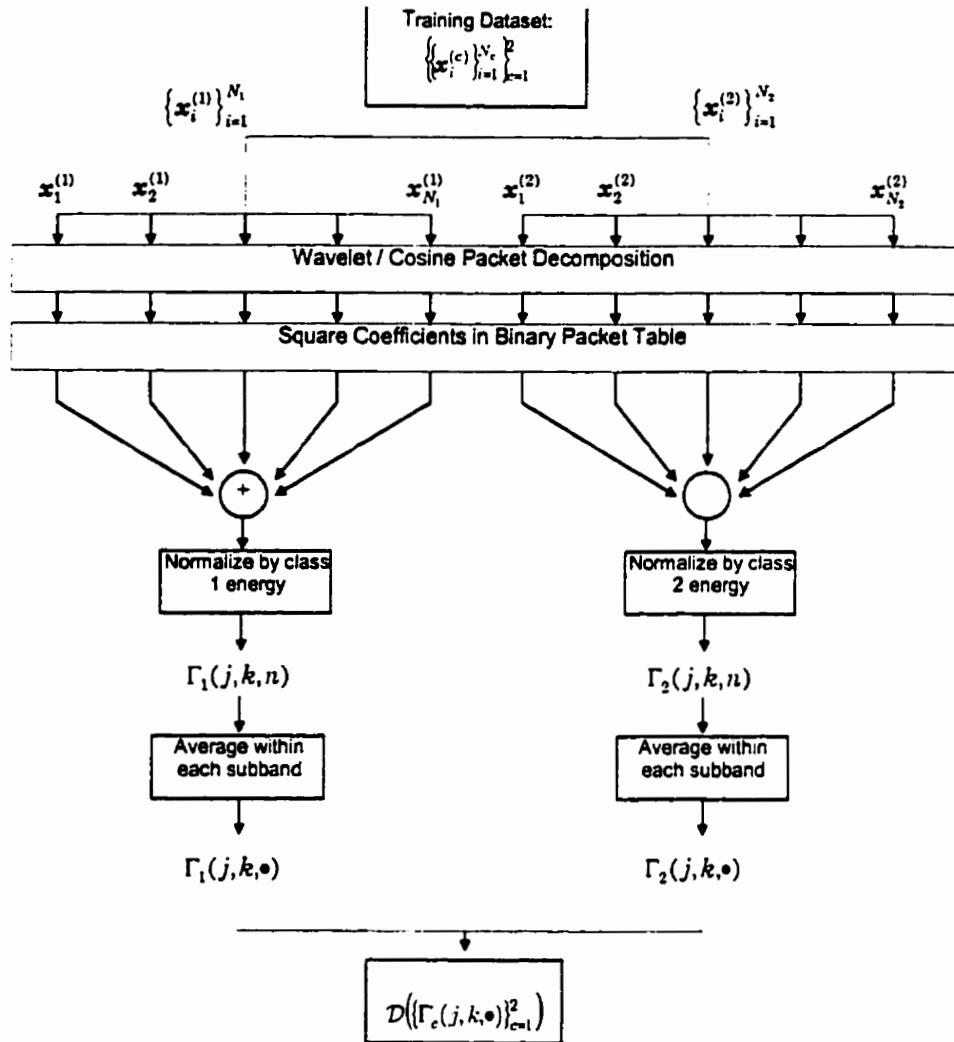


Figure 3.25 – Packet decomposition of a two-class training set into energy maps, suitable for computing the discriminant measure for the LDB algorithm.

Let $B_{j,k}$ denote a set of basis vectors belonging to the subspace $\Omega_{j,k}$, arranged in matrix form:

$$B_{j,k} = [w_{j,k,0}, w_{j,k,1}, \dots, w_{j,k,2^{n_0-j}-1}]^T. \quad (3.44)$$

Let $A_{j,k}$ represent the LDB for the training set restricted to the span of $B_{j,k}$, and let $\Delta_{j,k}$ be a work array containing the discriminant measure of the node (j,k) .

Algorithm 3.2 (The Local Discriminant Basis Algorithm) [Saito94].

Given a training dataset consisting of K classes of signals $\left\{ \left\{ \mathbf{x}_i^{(c)} \right\}_{i=1}^{N_c} \right\}_{c=1}^K$,

Step 0: Choose a time-frequency decomposition method. That is, specify a wavelet packet transform (i.e. a pair of QMFs) or a cosine packet transform. Specify the depth of decomposition J , and the discriminant measure \mathcal{D} .

Step 1: Construct the time-frequency energy maps Γ_c for $c = 1, \dots, K$.

Step 2: Begin at level J : set $A_{J,k} = B_{J,k}$ and $\Delta_{J,k} = \mathcal{D}\left(\{\Gamma_c(J,k,\bullet)\}_{c=1}^K\right)$ for $k = 0, \dots, 2^J - 1$.

Step 3: Determine the best subspace $A_{j,k}$ for $j = J-1, \dots, 0$, $k = 0, \dots, 2^j - 1$ by the following rule:

$$\text{Set } \Delta_{j,k} = \mathcal{D}\left(\{\Gamma_c(j,k,\bullet)\}_{c=1}^K\right)$$

$$\text{If } \Delta_{j,k} \geq \Delta_{j+1,2k} + \Delta_{j+1,2k+1},$$

$$\text{then } A_{j,k} = B_{j,k},$$

$$\text{else } A_{j,k} = A_{j+1,2k} + A_{j+1,2k+1} \text{ and set } \Delta_{j,k} = \Delta_{j+1,2k} + \Delta_{j+1,2k+1}.$$

Step 4: Order the N basis functions in the LDB by their power of discrimination (see below).

Step 5: Use the L ($\ll N$) most discriminating basis functions in the LDB for classifier features.

When Step 3 has been completed, we are left with $A_{0,0}$, which is the LDB restricted to the span of $B_{0,0} = \mathfrak{H}^N$: a complete orthogonal basis. The chosen LDB consists of a set of disjoint subspaces, which form a cover of the time-frequency plane. Each subspace $\Omega_{j,k}$ contains 2^{n_0-j} basis vectors. The total number of basis functions is always N , where $N = 2^{n_0}$ is the length of each signal $\mathbf{x}_i^{(c)}$. The

pruning algorithm is fast ($O(N)$) since the measure \mathcal{D} has been chosen to be additive.

Once the LDB has been selected, the N transform coefficients, each corresponding to a basis vector within the LDB, may be used as features for a classifier. It is desirable, however, to reduce the dimensionality of the representation for a classifier feature set. In Saito's algorithm this is done using *feature selection* methods, in steps 4 and 5. In Step 4, the basis functions in the LDB must be ranked to determine which are the most important for classification. As with the discriminant measure used for selecting amongst subbands to determine the LDB, a measure of class separability is used to assess the discriminant power of each basis function within the LDB, indexed by (j, k, n) :

$$\mathcal{D}(\Gamma_1(j, k, n), \dots, \Gamma_K(j, k, n)). \quad (3.45)$$

In Step 5, the dimension of the representation is reduced from N to L by keeping only the bases which provide the most discriminant information in terms of the time-frequency energy distributions between classes. Clearly, the choice L involves a classic tradeoff: decreasing L requires losing discriminatory information, but this helps counteract the curse of dimensionality. The best value of L depends upon the problem: the nature of the data and the type of classifier. In general, this can only be determined empirically.

Saito neglected to propose feature projection methods as an alternative to feature selection for the purpose of dimensionality reduction. Certainly, if the information tends to be dispersed throughout the time-frequency plane, it will be difficult to retain class separability information in a low-dimensional feature set using feature selection. Feature projection methods may prove to be superior in this situation,

as they seek to find the best combination of all features in a lower-dimensional projection. This will be investigated in Chapter 4.

3.4.2.3 Best Bases for Classification: Some Examples

A simple example is given here to bring clarity to the concept of the LDB algorithm. Consider again a chirp signal of length $N = 256$ sampled at 1000 Hz. A simple classification problem would be to distinguish between a “rising” chirp (increasing in frequency) and a “falling” chirp (decreasing in frequency). Now, increase the complexity of the problem a bit by adding some random noise to each signal with a signal to noise ratio (SNR) of 2 (roughly 6 dB), and consider a training set of 100 of these noisy chirps, 50 in each class. Four patterns from within each class of this dataset are shown in Figure 3.26.

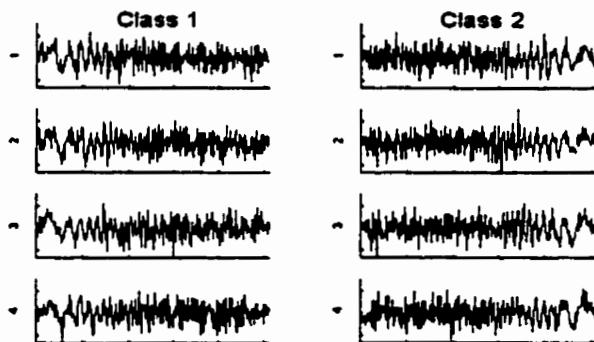


Figure 3.26 – Noisy chirp signals (SNR=2).
Class 1 are “rising” chirps, class 2 are “falling” chirps.

These training data were subject to WPT analysis using the LDB algorithm. The LDB was determined by pruning the packet tree according to a symmetric relative entropy discriminant measure. Figure 3.27(a) shows the tiling of the time-frequency plane generated by the LDB. Figure (b) shows the strength of the

discriminant measure for each basis vector within the LDB, evaluated using the same relative entropy measure as the pruning algorithm. The discriminant measure for each basis vector is displayed in the context of its time-frequency localization, bounded by its Heisenberg box.

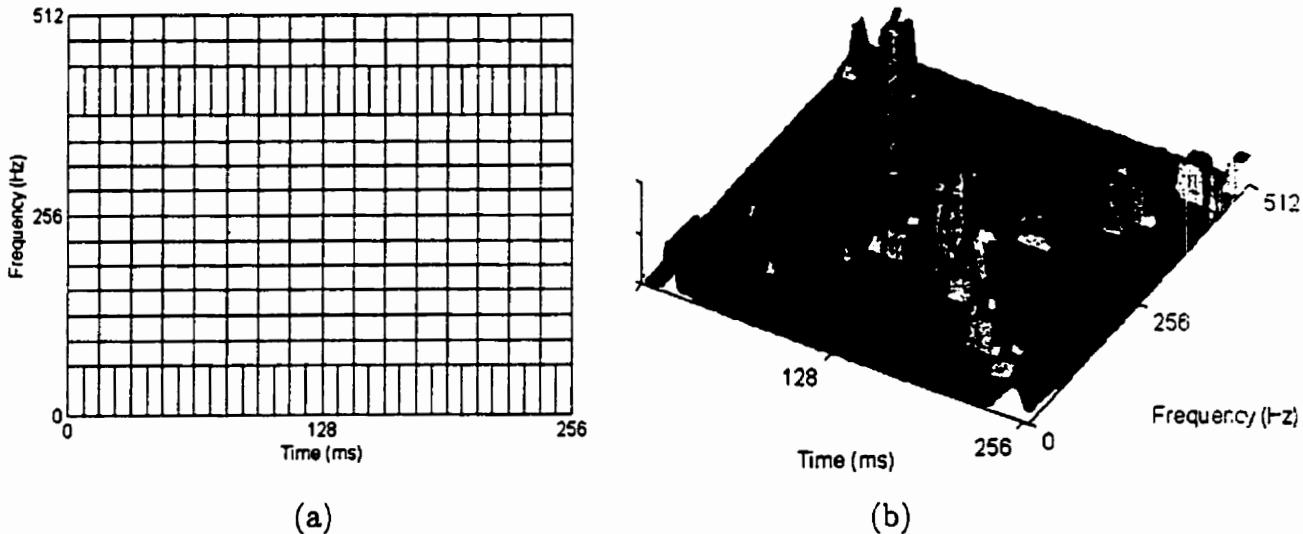


Figure 3.27 _ The properties of the WPT LDB. Figure (a) shows the tiling of the LDB, and (b) depicts the discriminant measure of each basis vector in the chosen LDB, bounded by its Heisenberg box in the time-frequency plane.

The most discriminant features are clearly along the diagonals. This is exactly what would be expected, since the noisy chirps differ most along the trajectories of their instantaneous frequencies. The tiling is fairly regular due to the linearly rising (and falling) frequency of the chirps.

The LDB consists of $N = 256$ basis vectors, and thus, 256 coefficients describe the decomposition of a given signal into the coordinate system specified by the LDB. According to the feature selection procedure of Saito's algorithm, the basis vectors must be ranked by their power of discrimination, and a suitable number of the most significant bases should be chosen as features for classification. Clearly, the basis vectors that should be chosen here are those that reside on the diagonals of

the time-frequency plane. Figure 3.28 shows a series of twelve two-dimensional scatterplots of the 24 WPT coefficients, ranked according to their discriminant power.

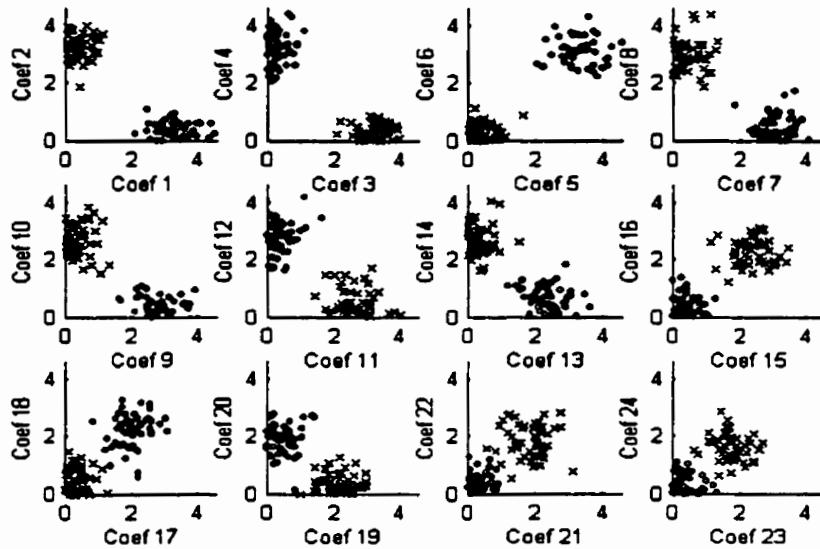


Figure 3.28 – A scatterplot of the coefficients of the 24 bases chosen from the LDB as feature extractors. Class 1 = x, class 2 = o.

Obviously, the discrimination problem here is trivial. With the SNR=2, the WPT LDB provides excellent clustering of the classes. Note that the coefficients are displayed according to their rank, and as the rank decreases, the class separation becomes less pronounced.

A more challenging classification task is provided with a SNR of 0.5 (roughly -6 dB). A feature set was again constructed from the most significant WPT LDB coefficients, ranked according to their discriminant power. With the knowledge that feature set dimension is a fundamental factor in classification (recall the discussion on the curse of dimensionality), it is important to know what dimension provides the best classification rate.

Figure 3.29 shows the classification error rate of the training set and test set data at all possible values of feature set dimension. A LDA classifier was used to generate these results.

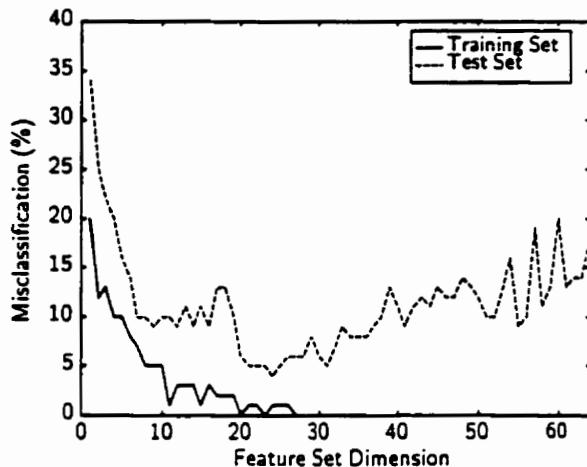


Figure 3.29 – Classification of noisy chirp signals (SNR=0.5) using a WPT LDB reduced using feature selection. The classification error using a LDA classifier is shown for a feature dimension ranging from one to 64.

As would be expected, both the training set and the test set classification rate is poor for a very low feature set dimension. The classification error of the training set becomes progressively lower with increasing dimension, since the LDA has more information upon which to draw linear discriminant bounds¹³. Added feature set dimension only aids classification of the test set up to a dimension of 25 however, beyond which the generalization suffers due to the curse of dimensionality. The procedure of selecting the optimum feature set dimension requires an independent dataset (a validation set). This is necessary for both feature selection and feature projection methods. The manner in which this is accomplished with transient MES datasets is described in Chapter 4.

¹³ The range of dimension is shown only for one to 64; for the remainder of the range up to N=256, the training set error is zero, and the test set error continues to inflate.

3.5 Summary

This chapter has provided the background necessary to develop a means of wavelet-based feature extraction for signal classification. The mathematics and concepts of wavelet theory have been introduced, and extended to include wavelet packet and cosine packet transforms. The key to the success of WPT and CPT representations is the selection of the *best basis* for the problem at hand. It has been shown that a basis can be determined so as to optimally localize discriminant information (i.e. a Local Discriminant Basis) by pruning a packet decomposition according to a class separability criterion. Some simple examples have demonstrated the advantages that WPT and CPT methods have over TFRs with fixed time-frequency tilings.

For the sake of completeness, the CPT has been presented as a orthogonal method to the WPT. It will not be included in further investigation however, because:

1. Preliminary work has shown that it does not match the performance of the WPT in transient MES classification, and
2. Its computational complexity is $O(N(\log_2 N)^2)$. This is somewhat larger than the STFT: $O(N \log_2 N)$, the WT: $O(N)$, and the WPT: $O(N \log_2 N)$.

In the next chapter, the performance of WT and WPT based feature sets will be evaluated in the context of the transient MES classification problem, and will be compared to the STFT and Hudgins' time domain feature set. This chapter will also compare the performance of feature projection based dimensionality reduction to the feature selection methods that have been used in previous work.

Chapter 4

Time-Frequency Methods for Myoelectric Signal Classification

This chapter provides an investigation of time-frequency methods and dimensionality reduction strategies that may be used to generate feature sets for classification of the transient myoelectric signal. The classification performance of each feature extraction / dimensionality reduction combination will be examined and explained with respect to the statistical and structural properties of the transient MES.

This chapter is organized as follows. Section 4.1 describes the data, the methods and the scope of this investigation. In the next four sections, a different feature set will be the subject of individual focus. In each case, the parameters of the feature extraction process will be empirically optimized for the transient MES classification problem, based upon a subject database acquired for this work. Correspondingly, the efficacy of dimensionality reduction strategies will be determined for each feature set. These feature extraction and dimensionality reduction strategies will be prescribed in the context of a statistical (LDA) and a neural (MLP) classifier. The feature sets under consideration are based upon time domain features (Section 4.2), the short-time Fourier transform (Section 4.3), the

wavelet transform (Section 4.4) and the wavelet packet transform (Section 4.5). The relative performance of each feature basis will be compared in Section 4.6, yielding a prescription of the best overall signal representation for transient MES classification.

4.1 Preliminary Issues

4.1.1 The Transient MES Data

It is the intention of this work to demonstrate the most effective means of classifying the transient MES signal accompanying the onset of motion of the upper limb. To provide a perfect representation of this problem would require an exceptionally large database of subjects. It is the experience of the author that significant differences exist in the patterns accompanying voluntary contraction amongst individuals. A reasonable encapsulation of the nature of this problem has been provided by a finite roster of subjects. The number of subjects was determined as a compromise between the time needed to acquire the data and the number needed to demonstrate the relative performance of different feature sets and dimensionality reduction techniques.

Two independent datasets were used in this study.

One Channel Data. This dataset comprises the data from thirteen subjects used in Hudgins' work. These data were collected using a single bipolar surface

electrode pair. An active electrode was placed over each of the biceps brachii and triceps brachii muscle groups, providing maximum pickup area of MES activity in the upper arm. The signals were acquired using a DAS16F A/D resident in a PC, and were sampled at 1000 Hz.

The subjects included nine normally-limbed and four limb-deficient individuals. Each subject was asked to produce four different types of anisometric contractions. All contractions began with the subject's arm by the side in a comfortable neutral position. The only constraint on the type of contraction was that the subject was asked to be consistent in reproducing the desired motion. A typical session had a subject producing 40 contractions for each of four classes of motion: elbow flexion, elbow extension, forearm pronation and forearm supination. Some subjects however, (primarily those with limb deficiencies) had difficulty reliably reproducing these motions, and some could not generate a full set of 40 contractions. Therefore, the data from some subjects includes a substituted motion (a co-contraction, for example) and a cardinality somewhat less than 160.

For the purpose of training and testing a classification system, the data were divided into a training set of 80 patterns and a test set of 80 patterns (typically). It has been determined in this work that 80 patterns is barely sufficient to properly train a classifier for this problem. When evaluating the test set performance, one would not want to have fewer than 80 patterns, as this would degrade the resolution of the error estimate. In the course of this investigation, it is necessary to have a validation set to select feature set parameters and specify the optimal

feature set dimension. Clearly, there are an insufficient number of patterns in Hudgins' data to provide for a validation set¹.

Two Channel Data. These data were collected specifically for this work. A roster of 16 normally-limbed subjects² provided data from two independent bipolar channels: one from the biceps brachii and one from the triceps brachii. The data were acquired using a handheld microprocessor-based A/D unit designed by the Institute of Biomedical Engineering at U.N.B. The data were serially uploaded from the battery-powered unit to a PC for permanent storage and further processing.

The same experimental protocol as described in Hudgins' work was used, except that 100 patterns were collected for each class, from each subject. Each subject produced a set of anisometric patterns corresponding to elbow flexion, elbow extension, forearm pronation and forearm supination. With 400 patterns available for each subject, it is possible to reserve some data for a validation set. One-hundred patterns were designated to the training set, 150 to the validation set, and 150 to the test set. Since 100 patterns are sufficient for the training process, a greater number of patterns were distributed to the validation and test sets to decrease the variance of the parameter estimation process.

¹ Cross-validation methods exist to allow validation data to exist with a limited number of exemplars. These methods however, add computational complexity to the analysis and are seldom as meaningful as an analysis based upon a truly independent validation set.

² It has been observed that limb-deficient individuals tend to have greater difficulty reproducing a set of target contractions than normally-limbed individuals. While this will affect the overall classification accuracy, there is no reason to suspect that this will affect the relative performance of different forms of signal representation. If differences of this type do exist, they are likely to be dependent upon many factors related to the nature of the limb deficiency which are too numerous to be identified here, and would certainly confound efforts to distinguish the relative efficacy of various feature sets. For this reason, any differences due to limb-deficiency were not of primary interest in this work.

These two-channel data will serve as the primary dataset for this work. All comparisons between feature sets and dimensionality reduction techniques will be based upon empirical measures from these data. The reasons for using these two-channel data are:

1. It has been shown that two-channel data yield superior classification results, as compared to one-channel data from the same set of contractions [Kurugati95]. Obviously, it is desirable to base a study of classification efficacy upon data of superior information content.
2. The number of patterns acquired permits the existence of a validation set of data. This is essential for systematic design of dimensionality reduction schemes.

4.1.2 Real-Time Considerations

The factor which imposes a real-time constraint upon the application of pattern recognition to prosthetic control is the system response time. It has been established that the delay between the initiation of a command by the operator and the desired actuation should be kept below 300 ms. The majority of this time will be spent collecting sufficient data to make reliable pattern recognition possible. Hudgins [Hudgins91] found that at least 200 ms of data was necessary, leaving at most 100 ms for signal processing. Hudgins did not consider transform-based features based upon the lack of an affordable microprocessor of sufficient computing capacity in 1991. This situation has changed dramatically: a DSP microprocessor of only modest capabilities today (costing only tens of dollars) can perform a 256-point FFT or DWT in a few milliseconds. This easily brings time-frequency transforms within the real-time constraints of prosthetic control.

4.1.3 Methodology

The objective of this investigation is to provide comparative evaluation of various forms of signal representation for the purpose of classifying transient MES patterns. The classification task is a multi-stage process however, and some of the stages are closely coupled. Consequently, the performance of each feature set must be set in the context of the chosen dimensionality reduction scheme and the chosen classifier, as illustrated in Figure 4.1.

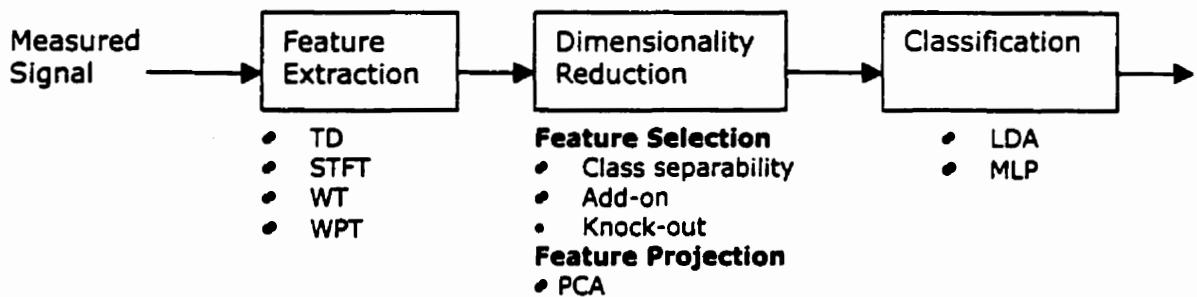


Figure 4.1 – The datasets, feature extraction methods, dimensionality reduction methods, and classifiers under investigation in this work.

This chapter will examine the performance of features based upon Hudgins' time domain (TD) statistics, the short-time Fourier transform (STFT), the wavelet transform (WT), and the wavelet packet transform (WPT). For each feature set, various methods of dimensionality reduction will be considered, including feature selection and feature projection. At the final stage of the classification task, two classifiers (LDA and MLP) with slightly different capabilities will be employed.

4.1.3.1 Datasets

The primary focus of this investigation is upon a dataset of two channel MES collected specifically for this investigation. The capabilities of feature sets, dimensionality reduction techniques, and classifiers will be investigated in detail

with respect to these data, which are intended to represent a generalization of the prosthetic control problem.

A “secondary” ensemble of datasets will be used: this is a roster of one-channel recordings collected in the same manner as the two channel data (these are the same data used by Hudgins). This data will be used in preliminary analyses involving time domain features, mainly to provide a context of Hudgins’ work within this investigation. It is not meant to serve as an indication of the relative performance of one versus two channel recordings, since these data were drawn from a different subject base, some of whom were limb-deficient individuals.

In Chapter 5, datasets composed of simulated MES patterns with an artificially imposed random structure will be considered. It will be shown that it is very difficult to simulate the intra-class energy dispersion in the time-frequency plane inherent in transient MES signals. This serves to demonstrate the structural complexity of the intra-class variance, and consequently, the difficulty of the transient MES classification problem.

4.1.3.2 Feature Sets

For each feature set, a number of design criteria exist that strongly influence its efficacy with respect to classification. For time domain features, the most important parameter is the means of segmenting the time record. When using a STFT basis, the type, size and degree of overlap of the window function is important. In wavelet analysis, the choice of mother wavelet is an influential parameter. Wavelet packet based methods depend not only upon the choice of

mother wavelet, but also upon the cost function used to determine the “best” basis for classification. At the beginning of each section which introduces a new feature set, the relevant design parameters will be investigated fully, allowing a specification which best suits classification of transient MES patterns.

4.1.3.3 Dimensionality Reduction

An appropriate scheme of dimensionality reduction can greatly improve the effectiveness of a given representation if it can preserve discriminant information while reducing the overall feature dimension. Dimensionality reduction is especially crucial when applying time-frequency transforms since the transform space is generally of very large dimension. The most effective means of dimensionality reduction depends upon the nature of the classification task. If the discriminant information tends to be concentrated within a small proportion of the features in the original feature space, *feature selection* methods will perform well. Although it is the role of the feature extraction stage to concentrate the discriminant information into as few variables as possible, highly unstructured data often yields a dispersion of information, regardless of the chosen representation. In this case, *feature projection* may serve the problem more suitably: the best linear combination of the original feature set may yield a much better feature set than the best subset of the original features. The dimensionality reduction scheme that provides the best complement to each feature extraction method will be determined.

The following describes the application of feature selection and feature projection in this investigation.

Feature Selection

An optimal feature selection scheme would consider all possible subsets of the original feature set. The importance of each candidate subset must be determined; ideally, importance should be conveyed by the generalization capabilities (test set probability of error) of a chosen classifier operating upon these data. This "ideal" feature selection scheme is not realizable, for two reasons. First, consideration of all possible subsets requires unreasonable computational expense, even for original feature sets of modest size. Second, a test set is generally not available to evaluate generalization; an estimate must be derived from the training set, such as probability of error³ or some class separability measure.

Three "suboptimal" schemes have been considered here:

1. *Class separability* (CS). Each feature is individually evaluated using a class separability measure. This method does not consider interactions between features, whether advantageous (multi-dimensional structures) or disadvantageous (correlation between features).
2. *Add-On* (AO). A probability of error estimate is used to determine the feature which, by itself, contributes most to training set discrimination. The "next best" feature is determined as that with the lowest probability of error when paired with the first. Features are "added-on" in this manner until the desired feature set dimension is attained.
3. *Knock-Out* (KO). This is basically the opposite of the add-on procedure. Beginning with the entire original feature set, the feature which contributes

³ The misclassification rate of the training set.

least to discrimination (the highest training set probability of error) is eliminated. This is iterated until the feature set is reduced to the desired dimension.

Each of these methods assigns an individual importance to each feature, which suggests that they may be *ranked*. An investigation of this ranking provides insight into the nature of the data, and allows a direct assessment of the best feature set dimension for generalization performance.

Feature Projection

A number of projection-based methods of dimensionality reduction were introduced in Chapter 2. The method that will be used to demonstrate the capabilities of feature projection in this work is *principal components analysis* (PCA). PCA seeks the coordinate system which best explains the variance in the data. As such, PCA's effectiveness as a dimensionality reduction method lies in its ability to model linear dependencies and to reject irrelevant information. Further, PCA has the advantages of having a closed-form solution (allowing a simple algebraic solution) and of automatically ranking the importance of the features in the projection space.

As an example, consider the STFT coefficients of a dataset of transient MES patterns. The dataset consists of four classes of patterns, each containing 20 exemplars. The STFT of each pattern of length $N = 256$ samples contains 131 coefficients, each corresponding to a time-frequency cell. First, the STFT coefficients were ranked according to the CS feature selection. Figure 4.2 shows

the top 32 of these ranked coefficients, plotted as a series of two-dimensional scatterplots.

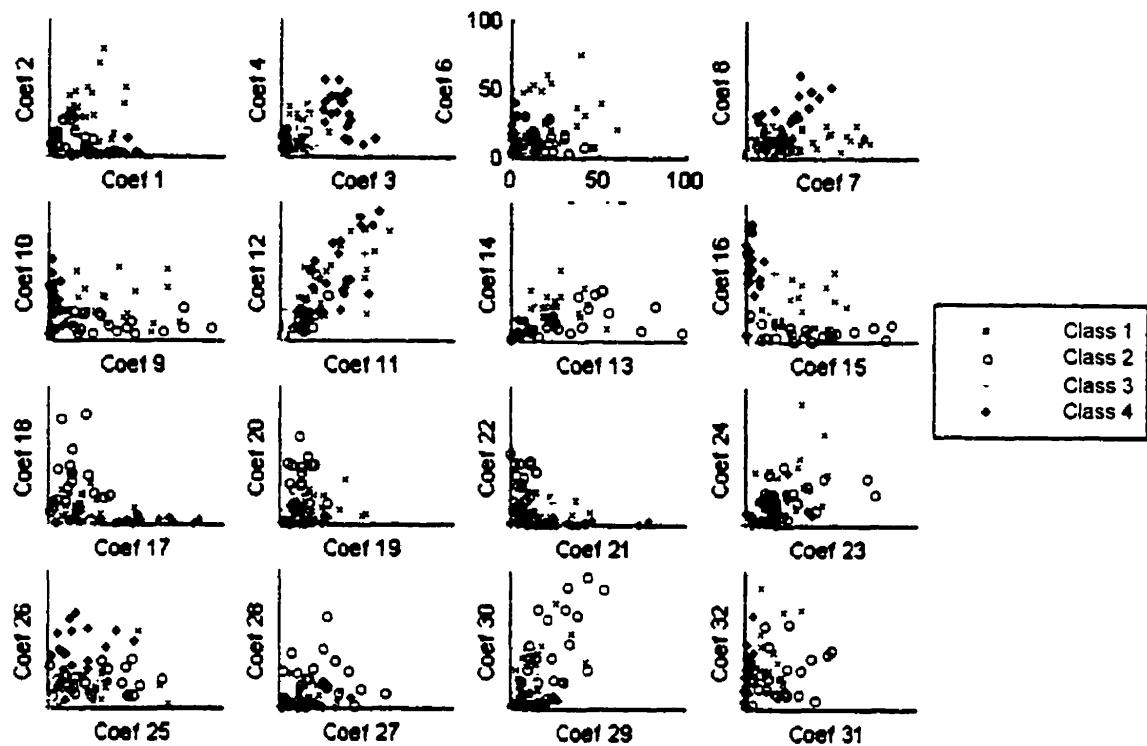


Figure 4.2 – A series of scatterplots of the CS-ranked STFT coefficients of a transient MES dataset. The pairwise scatterplots are presented such that the most highly-ranked coefficients are shown first (top left), and the least separable coefficients are shown last (bottom right).

The STFT data are extremely variable, allowing little distinction amongst the classes. This is because the discriminant information is dispersed throughout the time-frequency plane. Due to the high dimension of the STFT (131), each coefficient possesses a substantial degree of intra-class variance. Undoubtedly, significant linear dependencies exist amongst the STFT coefficients as well. For a classifier to perform well, the discriminant information must be concentrated into as few coefficients as possible.

The STFT coefficients were then subject to PCA. The top 32 projected features (ranked according to their eigenvalues) are shown in Figure 4.3.

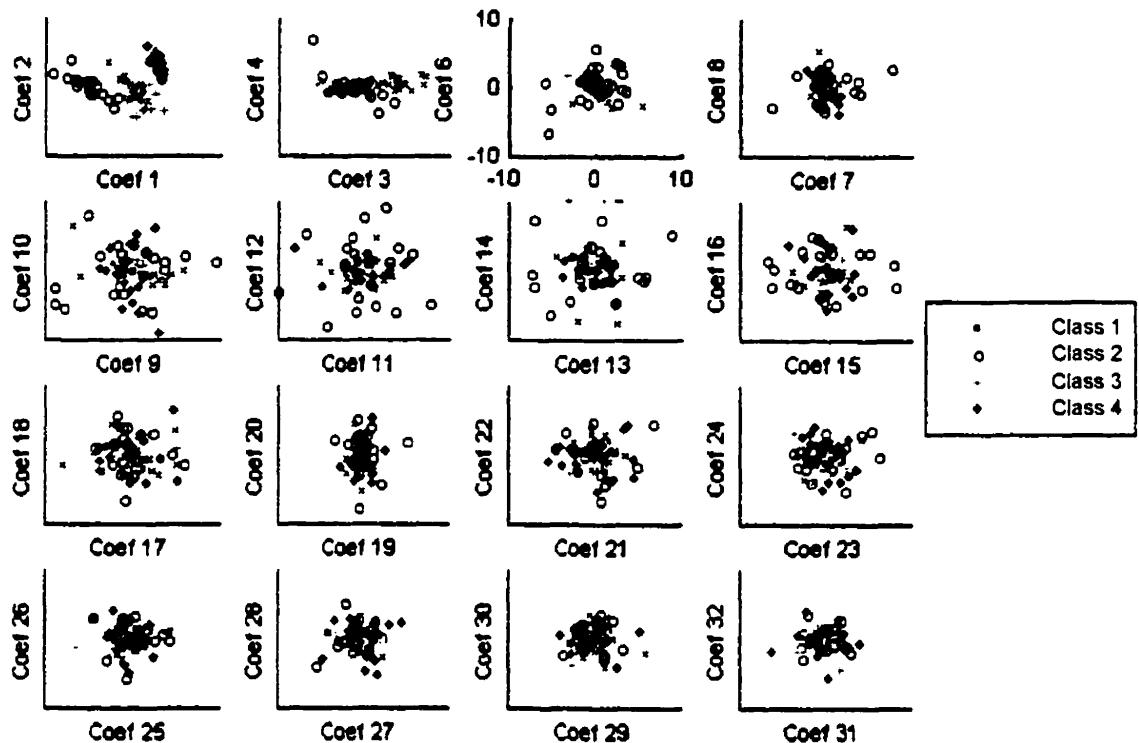


Figure 4.3 – A series of scatterplots of the PCA-ranked STFT coefficients of a transient MES dataset.

The ability of PCA to model linear dependencies and to discard irrelevant information is evident, as most of the discriminant information now appears to be embedded in the first three or four PCA features. PCA will be shown to be an extremely effective technique for reducing the dimension of the TFR-based feature sets in this chapter; as each TFR yields dispersed information in a high-dimensional feature space.

Other methods exist that offer more complex objective functions such as projection pursuit [Huber85] and independent component analysis [Karhunen97]. Other

projection methods allow nonlinear projections, such as nonlinear PCA [Kramer91] and Kohonen maps [Kohonen89]. These techniques however, require exploratory analysis or nonlinear optimization methods, which may be inexact or extremely computationally intense. PCA has been chosen as the representative method here due to its straightforward interpretation and its relatively moderate computational burden.

4.1.3.4 Classifiers

Although the focus of this work is not upon the classification stage but upon the signal representation, there are some important observations to be made with respect to the capabilities of the classifier. The representative classifiers chosen here are the LDA and the MLP: these are simple, but effective representatives of *statistical* and *neural* classifiers.

Most statistical classifiers can be interpreted to follow the Bayesian classification principle. They either explicitly estimate the class densities and *a priori* probabilities, or the optimal discriminant functions directly by regression. They are therefore defined by explicit error functions operating globally upon the data. Computational requirements can be extensive, depending upon the dimension of the data. The LDA is perhaps the simplest and most widely used statistical classifier. It compromises, but vastly simplifies the problem of estimating the class-conditional densities in the Bayes classifier by modeling them as multivariate normals. Depending on whether unequal or equal class covariances are assumed, a discriminant surface results that is either quadratic (QDA) or linear (LDA), respectively.

Neural algorithms on the other hand, are on-line learning systems, intrinsically nonparametric and model-free. The computational aspect of neural networks is central: learning should be accomplished by many simple, local computations using the available input and output signals (as do real neurons). No heavy numerical algorithms should be required, such as matrix inversions. The knowledge learned from examples is stored in a distributed manner amongst the connection weights. As a result, neural networks are flexible in their architecture, and are capable of constructing arbitrarily complex decision boundaries. The greatest challenge in implementing neural network classifiers is in tuning them to achieve the best possible generalization performance. This has to do with network size, the training algorithms, and augmentative strategies, such as weight decay. The MLP is the simplest, most studied, and most widely-used of all neural network classifiers. For this reason, it has been chosen as the “neural” representative here.

The following are some remarks pertaining to the application of a LDA and a MLP to the transient MES classification problem:

1. Although the LDA can be computationally expensive, it is not so for the relatively low feature dimensions that will be encountered in this work.
2. Whereas the LDA provides a deterministic solution, the MLP provides a stochastic solution. This is due to the fact that (i) a random starting point is used, (ii) the learning algorithm is stochastic, (iii) the stopping criteria is heuristic.
3. The number of hidden layer nodes in the MLP which yields the best generalization performance was determined from a validation dataset (see the next section). The number of hidden layer nodes was varied from four to ten for each feature set (TD, STFT, WT, WPT). Although the validation set

classification error was relatively insensitive to the number of hidden layer nodes over this range, a size of eight was found to provide the best generalization, on average, across all subjects. This was consistent for each feature set.

4. The MLP stopping criterion used here was a fixed number of training epochs (200), based upon observation of the training dynamics of the validation set data. For most subjects, the validation set classification error nears a minimum at 200 epochs, with little incidence of overtraining.

4.1.4 Dimensionality Specification *via* Validation

The design of the signal representation and the classifier should be strictly based on the *training set* only. The separate *test set* can be employed to obtain an unbiased estimate of the performance of the system thus designed. In order to select design parameters with the intent of maximizing generalization performance, some of the training set must held back to form a *validation set*⁴. This validation set is used in the same manner as a test set, allowing an approximation of the generalization ability. The number of patterns in each dataset (for each of four classes) was chosen to be sufficiently large to meet the following requirements.

⁴ To utilize the available training sample efficiently, cross-validation can be used. In v -fold cross-validation the training sample is divided into v disjoint subsets. One subset at a time is held back, a classifier is designed based on the union of the remaining $v-1$ subsets, and then tested using the subset held back. Cross-validation approximates the design of a classifier using all of the available training data and then testing it on an independent set of data [Holmstrom97]. The need for cross-validation methods exists only if there are insufficient data to reserve an exclusive validation dataset. Clearly, it is more straightforward to proceed with one validation set, rather than resorting to cross-validation methods. This was recognized prior to undertaking data collection, and sufficient data were acquired to have an independent validation set.

The training set must be large enough to fully train the classifier. For both the LDA and MLP classifiers, it was empirically determined that a training set size of 100 patterns was sufficient, for most subjects.

The test set must be large enough to provide sufficient resolution in estimating the (percentage) classification error. For two-channel datasets, the percentage error is often below 10%. To provide an adequate comparison amongst methods, a resolution of less than 1% would be preferred. Therefore, 150 patterns were allotted to the test set, subtending a resolution of $100/150=0.67\%$.

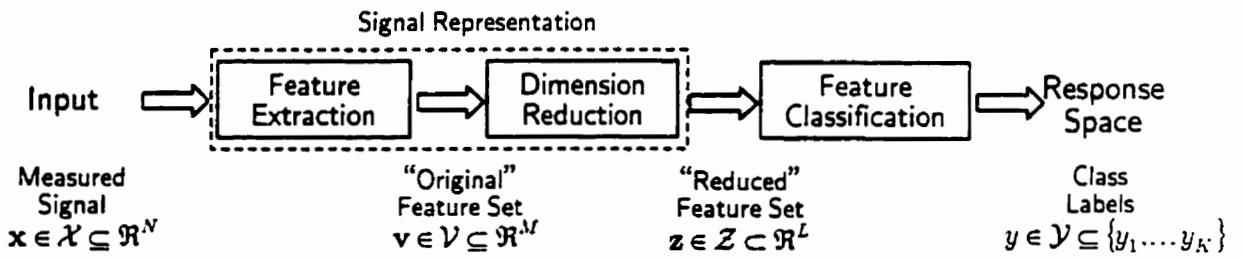
The validation set must be sufficiently large to provide a good estimate of test set classification performance. The validation set provides an estimate of the generalization performance of the feature set / classifier combination. Assuming that the validation set patterns are drawn from the same probability distribution as the test set patterns, a larger validation set will decrease the statistical variance when estimating the test set generalization performance. The number of patterns required to provide a stable estimate is somewhat nebulous: it was empirically determined that increasing the validation set size beyond that of the test set did not offer much improvement. Therefore, 150 patterns were designated to the validation set.

In total, $100+150+150=400$ patterns were collected from each subject (100 in each of four movement classes).

In this investigation, the validation set was used to provide specification of feature extraction parameters, feature set dimension, and classifier architecture. In the

sections that follow, the feature extraction parameters of each feature set are optimized by performance analysis of the validation set. As mentioned previously, the best MLP hidden layer size was determined in the same manner.

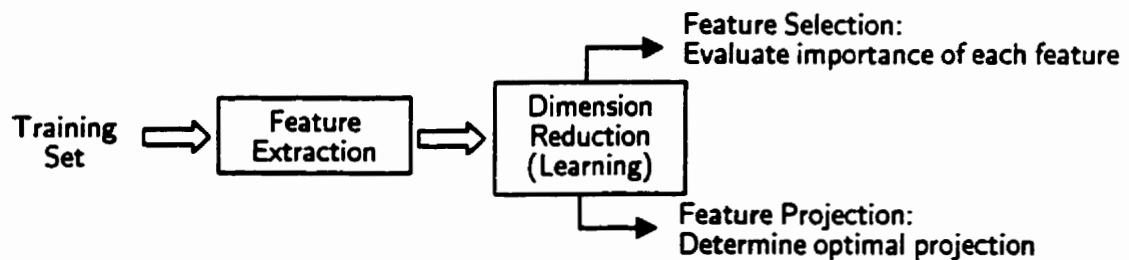
The validation set is used universally in this work to specify the “optimal” dimension of a reduced feature set when using either feature selection or feature projection. Recall the block diagram that describes the signal subspaces at each stage of the classification process.



The following algorithm describes the process of estimating feature set dimensionality using a validation set.

Step 1: Determine the dimensionality reduction parameters.

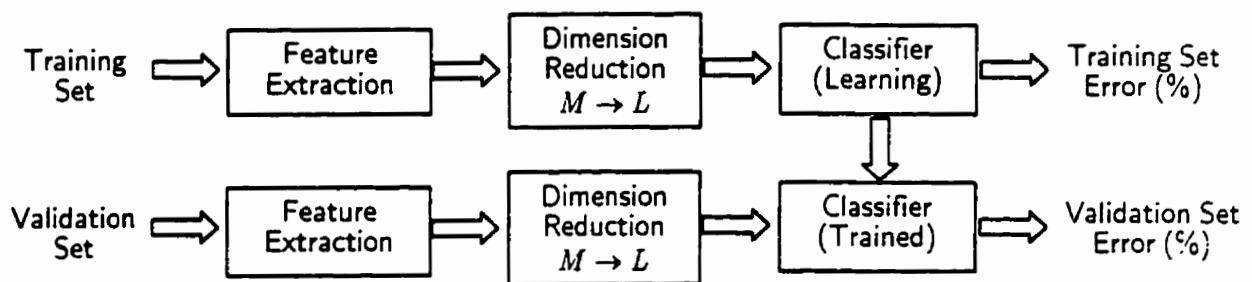
- (i) Extract features from the training set.
- (ii) Compile a set of features, ranked according to importance:
 - (a) Feature Selection: evaluate the importance of each feature by some measure of class separability or probability of error.
 - (b) Feature Projection: determine the optimal linear projection of the features, project the features, and rank the projected features.



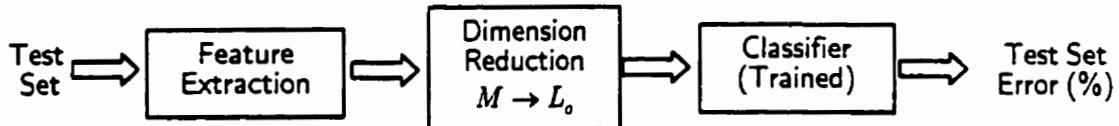
Step 2: Determine the optimal feature dimension via validation

- (i) Extract features from training / validation sets.
- (ii) Perform dimensionality reduction to produce reduced feature sets of the L most important features, for $L_{\min} \leq L \leq L_{\max}$.
- (iii) Train classifier (via training set) for each $L_{\min} \leq L \leq L_{\max}$.
- (iv) Classify reduced feature sets of validation data for all $L_{\min} \leq L \leq L_{\max}$.
- (v) Determine the L that minimizes the validation set error ($L = L_o$).

This approximates the reduced feature dimension that minimizes the test set (generalization) error.



Step 3: Evaluate the test set error at the optimal feature set dimension L_o .



A typical response of the training, validation, and test set error rate as a function of feature dimension is shown in Figure 4.4.

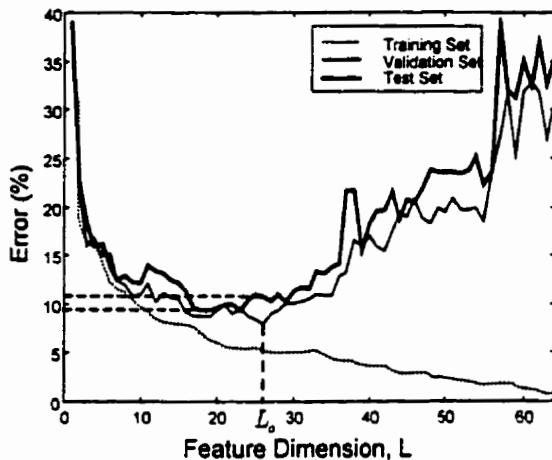


Figure 4.4 – An example of validation based dimensionality specification. The classification error of the training set, the validation set and the test set are shown at all possible values of feature set dimension. The dimension L_0 , determined from the validation set, is used to evaluate the test set performance.

The training set error declines with feature dimension, as a classifier (of sufficient capacity) can always improve its discriminant ability with this added information. The validation and test sets exhibit an “optimal” dimension; at this point, added feature dimensions would degrade generalization performance due to the curse of dimensionality. Clearly, the validation set error provides an *estimate* of the true generalization performance (the test set error), and therefore the estimate of this optimal dimension, L_0 , is an estimate of the optimal dimension for the test set. In general, L_0 is not the best dimension for the test set; this is evident in the difference between the test set error at L_0 and the minimum test error.

4.1.5 Summary

This section has provided an overview of the methods that will be used to investigate the performance of time-frequency based feature sets. The following sections provide a detailed analysis of each the TD, STFT, WT and WPT feature sets. For each feature set, the optimal configuration for transient MES classification is determined empirically by validation, and the benefits of dimensionality reduction are examined. These analyses lead to a prescription of the best combination of feature extraction and dimensionality reduction for the transient MES classification problem.

4.2 Time Domain Features

Hudgins' [Hudgins93] discovery of a deterministic component in the transient MES waveform prompted a new challenge: how can this temporal structure be captured in a feature set? Hudgins' primary criteria were twofold: the features must provide good class separability (to provide classification accuracy), and they must require minimal computational effort (to reduce delays in state selection).

4.2.1 Derivation of the Time Domain Feature Set

Although the transient MES patterns have some loose structure in the temporal waveform (a deterministic component), there is a great deal of intra-class variability (a random component). An attempt to classify these patterns using the sampled waveforms would result in very poor classification performance. The high variability of the features, and the high dimension of the input space would yield a sparsely populated feature space: no classifier could be expected to generalize well under such extreme effects of the curse of dimensionality. If one were to compute statistics based upon the entire record however, the temporal structure would be lost. The approach taken then, was to segment the transient waveform and compile a feature set based upon statistics from each segment. This is illustrated in Figure 4.5.

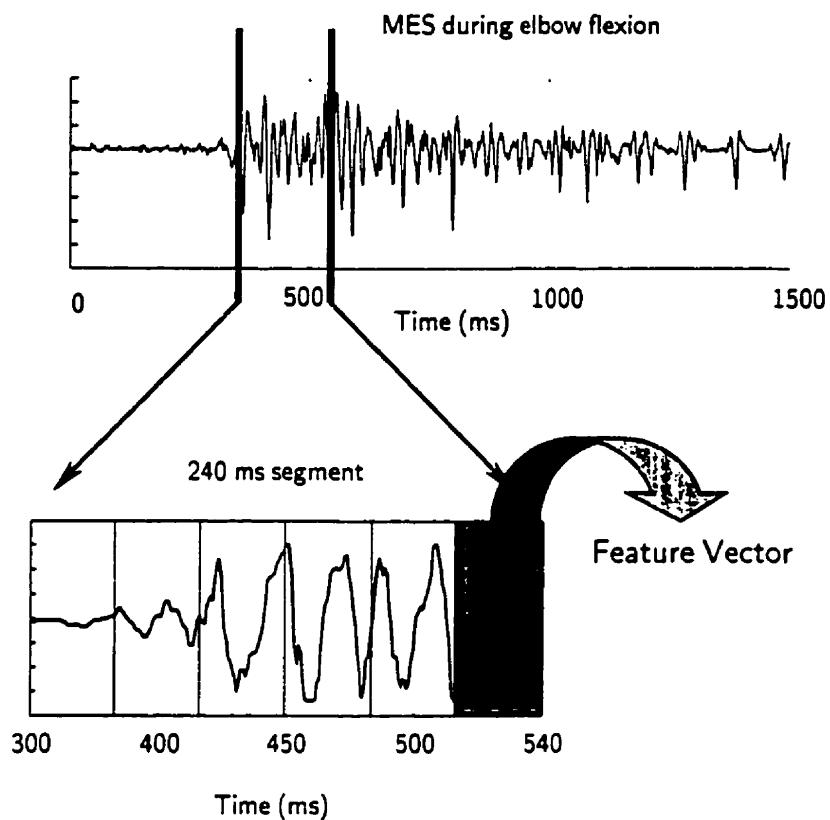


Figure 4.5 – Waveform segmentation for time domain feature extraction.

Hudgins experimented with different segment lengths in an attempt to minimize the classification error. For highly structured patterns, increasing the number of segments will increase the dimension of the feature set and therefore, the amount of information available to the classifier. For patterns with little temporal structure, smaller time segments will result in larger feature variance, reducing the relative class separability and degrading classifier accuracy. Based upon empirical results, Hudgins decided upon a scheme of five 40 ms segments, plus an extra “segment” comprising the mean value of the features computed over all six segments. The following features were chosen to represent the MES patterns:

Mean Absolute Value (MAV) – An estimate of the mean absolute value of the signal \mathbf{x} in segment i which is L samples in length is given by

$$\bar{x}_i = \frac{1}{L} \sum_{k=1}^L |x_k| \quad \text{for } i = 1, \dots, I \quad (4.1)$$

where x_k is the k^{th} sample in segment i and I is the total number of segments in the record.

Mean Absolute Value Slope (MAVS) – This is simply the difference between the sums in adjacent segments, i and $i+1$, as defined by

$$\Delta \bar{x}_i = \bar{x}_{i+1} - \bar{x}_i \quad \text{for } i = 1, \dots, I. \quad (4.2)$$

Zero Crossings (ZC) – A simple frequency measure can be obtained by counting the number of times the waveform crosses zero. A threshold (ε) must be included in the zero crossing calculation to reduce the noise induced zero crossings. Given two consecutive samples x_k and x_{k+1} , increment the zero crossing count, ZC, if

$$\begin{aligned} & \{x_k > 0 \text{ and } x_{k+1} < 0\} \text{ or } \{x_k < 0 \text{ and } x_{k+1} > 0\} \\ & \quad \text{and} \\ & \quad |x_k - x_{k+1}| \geq \varepsilon \end{aligned} \quad (4.3)$$

Slope Sign Changes (SSC) – A feature which may provide another measure of frequency content is the number of times the slope changes sign. Again, a suitable threshold must be chosen to reduce noise induced slope sign changes. Given three consecutive samples, x_{k-1} , x_k and x_{k+1} , the slope sign change, SC, is incremented if

$$\begin{aligned} & \{x_k > x_{k-1} \text{ and } x_k > x_{k+1}\} \text{ or } \{x_k < x_{k-1} \text{ and } x_k < x_{k+1}\} \\ & \quad \text{and} \\ & |x_k - x_{k+1}| \geq \varepsilon \text{ or } |x_k - x_{k-1}| \geq \varepsilon \end{aligned} . \quad (4.4)$$

Waveform Length (WL) – A feature that provides information on the waveform complexity in each segment is the waveform length. This is simply the cumulative length of the waveform over the segment, defined as:

$$l_0 = \sum_{k=1}^L |\Delta x_k|, \quad (4.5)$$

where $\Delta x_k = x_k - x_{k-1}$. The resultant values indicate a measure of waveform amplitude, frequency, and duration all within a single parameter.

The composite of these features from each segment forms the feature vector; by segmenting the waveform into I frames one specifies a feature set dimension of $(I + 1)$ segments $\times F$ features/segment (the extra “segment” includes the mean features). In Hudgins’ analysis, five segments were used, yielding a feature set dimension of 30. When processing two-channel data, each channel is subject to segmentation and feature extraction independently, and an aggregate feature set formed by concatenation. For example, if five 40ms segments were used on each channel, the resulting feature set would consist of $(5 + 1)$ segments $\times 5$ features/segment $\times 2$ channels = 60 features .

Upon inspection of the transient MES patterns, it is evident that the deterministic structure is much more pronounced in certain contraction types. Other types present an almost random nature from the onset. Although the variance in the time structure of these signals is high, the waveform statistics may be stable enough to allow adequate pattern classification. Hudgins sought to determine

which of these features contributed significantly to the classification task. An ANOVA was used to determine the ratio of feature variance between classes to within classes. Using this metric, almost all features (in almost all segments) showed significant differences between the classes, for each of ten subjects. This simple variance ratio does not fully describe the value of a feature's inclusion in a feature set. It will be shown that systematic methods of dimensionality reduction can improve classifier generalization. Choosing either the best subset (*feature selection*) or the best linear combination (*feature projection*) of the time domain features can simplify the task of the classifier, and curtail the effects of the curse of dimensionality.

4.2.2 Feature Set Parameters

If a time domain feature set is to be used as a signal representation for classification, it should be optimized with respect to this application. Further, it should be optimized with respect to the data of interest. Two datasets are of interest here:

1. A roster of one channel data from 13 subjects, drawn from Hudgins' work. These data will allow the present analysis to corroborate Hudgins' observations, and provide further perspective on parameter selection when using one channel data.
2. A roster of two channel data from 16 subjects, collected during the course of this work. This dataset will be the primary focus of this work; the selection of the best feature extraction parameters will be used to compare the performance of time domain features with time-frequency based features.

The design parameters relevant to time domain feature extraction are

- i) *Record length*, and
- ii) *Waveform segmentation*

4.2.2.1 Record Length

The most important parameter in time domain feature extraction is record length. If the record length is too small, insufficient information is present in the data to allow accurate pattern classification. On the other hand, the record length cannot be arbitrarily large, since classification must be performed in less than 300ms from the onset of contraction. Using a sampling rate of $f_s = 1000 \text{ Hz}$ (1 sample per millisecond), Hudgins determined that 200ms was the longest record length that would allow sufficient time (<100ms) to compute the time domain feature set and perform classification. This constraint will be relaxed somewhat due to the availability of greater computational capacity.

One Channel Data

For each of 13 subjects in Hudgins' original dataset, the test set classification error was computed using time domain feature sets derived a variable number of 40ms segments. The number of segments ranged from one to seven, specifying a record lengths ranging from 40 to 280 ms. Figure 4.6 illustrates the test set classification performance.

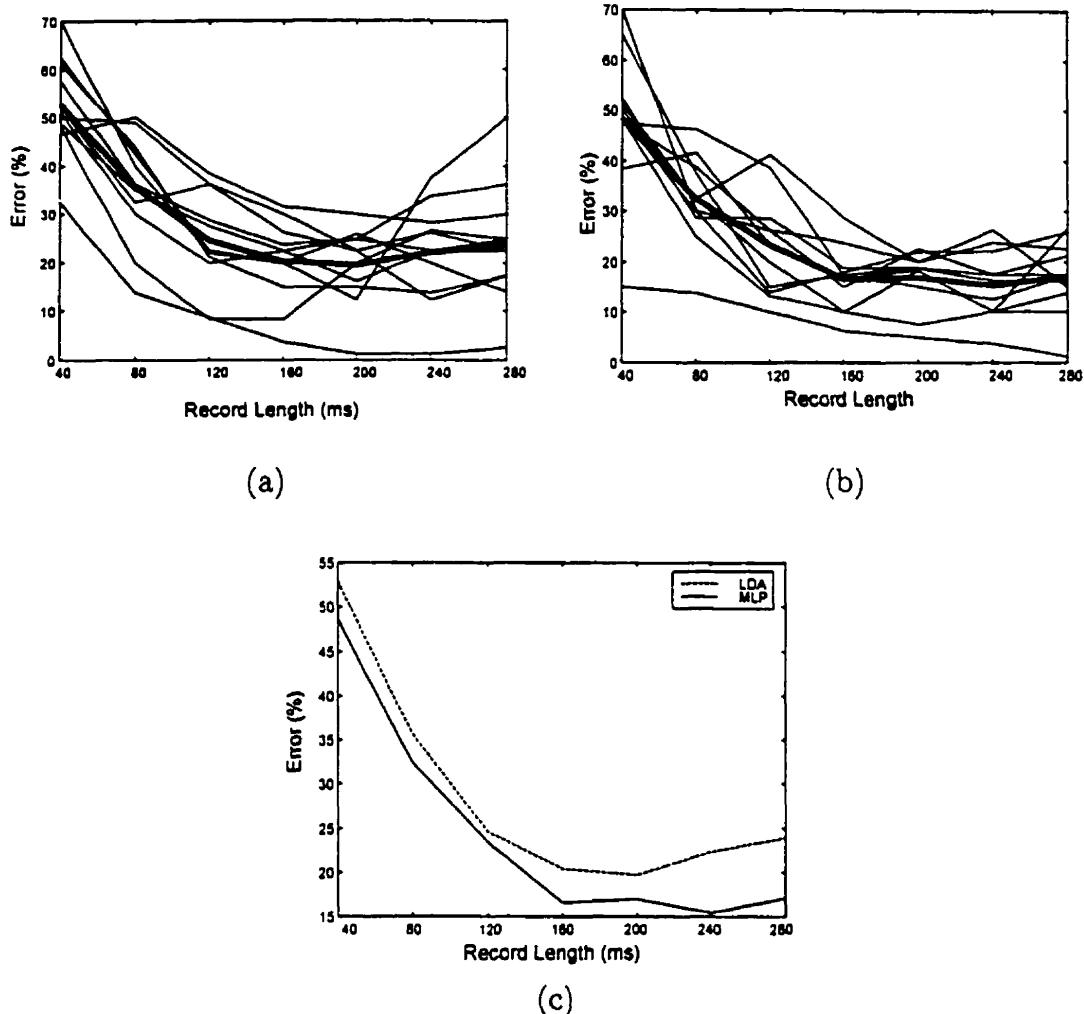


Figure 4.6 – Hudgins' one channel dataset: the effect of record length upon the test set classification rate. Figure (a) depicts the LDA error rate for each subject, superimposed by the mean across all subjects (shown as the heavy line). Figure (b) shows the same using a MLP classifier. Figure (c) compares the mean error rate of the LDA and the MLP.

Figure 4.6 (a) and (b) depict the classification performance of each subject using LDA and MLP classifiers, respectively. Superimposed on each plot is the mean classification rate across all subjects. Figure 4.6 (c) provides a direct comparison between the mean response of the LDA and MLP. The LDA seems to perform best using five segments (200 ms), while the MLP reaches its lowest error rate at six segments (240 ms).

Two Channel Data

The same analysis of record length effect was performed upon all subjects in the two channel dataset. The averaged test set classification error is shown in Figure 4.7.

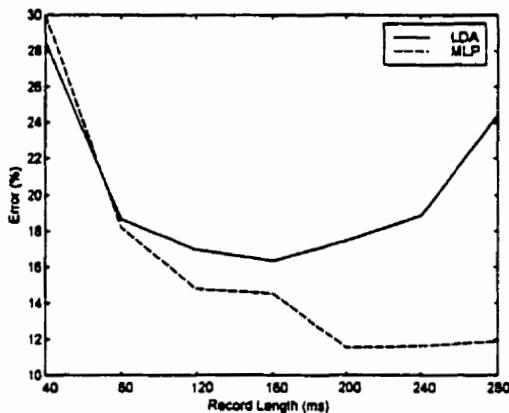


Figure 4.7 ... Two channel dataset: the effect of record length upon the test set classification rate, averaged across all subjects.

This two channel data exhibits the same trends as the one channel data. The LDA performance degrades as the number of segments grows larger than four (a record length of 160ms), and the MLP seems to perform best with five segments (record length equals 200ms). These results are somewhat misleading, however. They seem to suggest that if one were to use an LDA classifier, a very short time record is preferable. One must appreciate that as the number of segments I increases, the dimensionality of the resulting feature set grows as $(I+1)$ segments $\times F$ features/segment $\times \#$ channels. For the one channel data we have $(I+1) \times 5$ dimensions; for the two channels data we have $(I+1) \times 5 \times 2$ dimensions. Due to a finite training set size (80 patterns for the one channel data, 100 patterns for the two channel data), the curse of dimensionality will adversely affect generalization performance as the feature set dimension grows large. The LDA is particularly vulnerable to the effects of the curse of dimensionality, as is evident in

the figures above. The MLP can incorporate additional feature dimensions (even if they contain irrelevant information) without as dramatic an effect.

Longer record lengths undeniably contain more information, but this information is useful only if it does not overwhelm the classifier. The performance with respect to record length will certainly differ if the feature set is subject to some form of dimensionality reduction. Indeed, it will be shown in a later section that with an appropriate form of dimensionality reduction, longer record lengths do yield better classification performance. The record length will be limited to 240 ms however, in order to meet the real-time constraints of the classifier.

4.2.2.2 Waveform Segmentation

The other parameter in time domain feature extraction is the *segmentation scheme*. Using a few chosen subjects as benchmarks, Hudgins determined that five 40 ms segments provided the best segmentation of a 200 ms record, using a MLP classifier. A somewhat more comprehensive approach has been taken here. The evaluation of each segmentation scheme has been done for all subjects, using both LDA and MLP classifiers. Assuming a maximum record length of 240 ms, many possible combinations of segment length (L) and the number of segments (I) were considered. A maximum record length of 240 samples allows many integer-valued (L, I) pairs, assuming non-overlapping segments. Table 4.1 delineates the segmentation schemes used, and the corresponding record length.

| Number of Frames, I | Segment Length, L (ms) | | | | | | | |
|-----------------------|--------------------------|-----|-----|-----|-----|-----|-------|-------|
| | 10 | 20 | 30 | 40 | 60 | 80 | 120 | 240 |
| 1 | 10 | 20 | 30 | 40 | 60 | 80 | 120 | 240 |
| 2 | 20 | 40 | 60 | 80 | 120 | 160 | 240 | 480 |
| 3 | 30 | 60 | 90 | 120 | 180 | 240 | 360 | 720 |
| 4 | 40 | 80 | 120 | 160 | 240 | 320 | 480 | 960 |
| 6 | 60 | 120 | 180 | 240 | 360 | 480 | 720 | 1,140 |
| 8 | 80 | 160 | 240 | 320 | 480 | 640 | 960 | 1,920 |
| 12 | 120 | 240 | 360 | 480 | 720 | 960 | 1,140 | 2,880 |

Table 4.1 – The quantized ranges of segment length (L) and the number of frames (I) used as possible waveform segmentation schemes. The table entry indicates the total record length for each (L, I) pair. Only record lengths less than 240 ms (the light gray region) or equal to 240 ms (the white region) were considered.

In Figure 4.8 the classification error of the one channel dataset is illustrated as a vertical bar graph, using each of the segmentation schemes in the table. The plots depict the training set and test set error averaged across all subjects, using a MLP classifier.

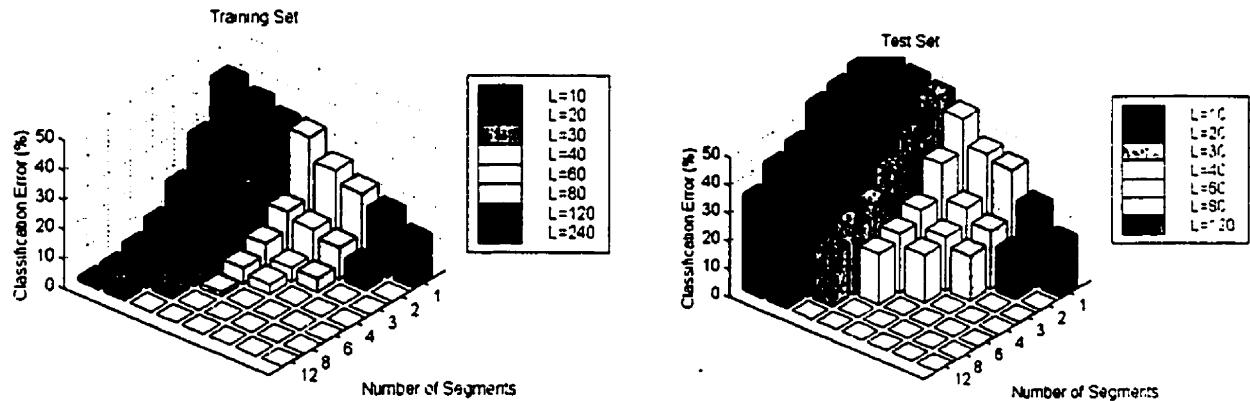


Figure 4.8 – One channel data: the classification error upon the training set (left) and the test set (right) using a MLP classifier, using a variety of segmentation schemes. The rates shown are the average across all subjects.

Clearly, these plots could be shown for each subject, but this amount of information would be excessive. The average across subjects allows visualization of trends made apparent by averaging, and implies some generalizations. In these bar graphs, the classification rate is depicted as a function of segment length (L)

and the number of segments (I). For all segmentation schemes, the segments are adjacent and non-overlapping, so that the full record length is $N = L \times I$. It is evident in both the training set and the test set that the lowest classification error is achieved when using the greatest allowable record length ($N=240\text{ms}$). The locus of $N=240$ is the front diagonal of the bar charts. The training set exhibits increasing classification performance as the number of segments increases. This is to be expected, since a greater number of segments subtend a larger feature set dimension. Increasing feature set dimensionality usually allows greater discriminant ability in a training set, as long as the classifier is of sufficient capacity.

Larger feature set dimension only helps the test set up to a certain point, however. The additional information afforded by more features is, at some point, overwhelmed by the need to construct discriminant bounds in a higher dimensional space. The effects of the curse of dimensionality are evident in the test set error rates. Segmentation schemes with many frames experience poor generalization performance. Since the relevant measure of performance is the ability to generalize, the best segmentation scheme must be chosen on this basis.

This analysis, repeated using a LDA classifier, shows very similar trends. The main difference is that the test set error experiences greater degradation when the number of segments grows too large. This segmentation analysis performed upon the two channel dataset generates results that are very similar, for both the LDA and MLP classifiers.

If we restrict our attention to the locus of $N=240$, the information generated by this segmentation analysis can be interpreted more easily. Figure 4.9 shows the

test set classification rate averaged across all subjects as a function of the number of segments I (or segment length $L = 240/I$). The results are shown for both the one and two channel datasets.

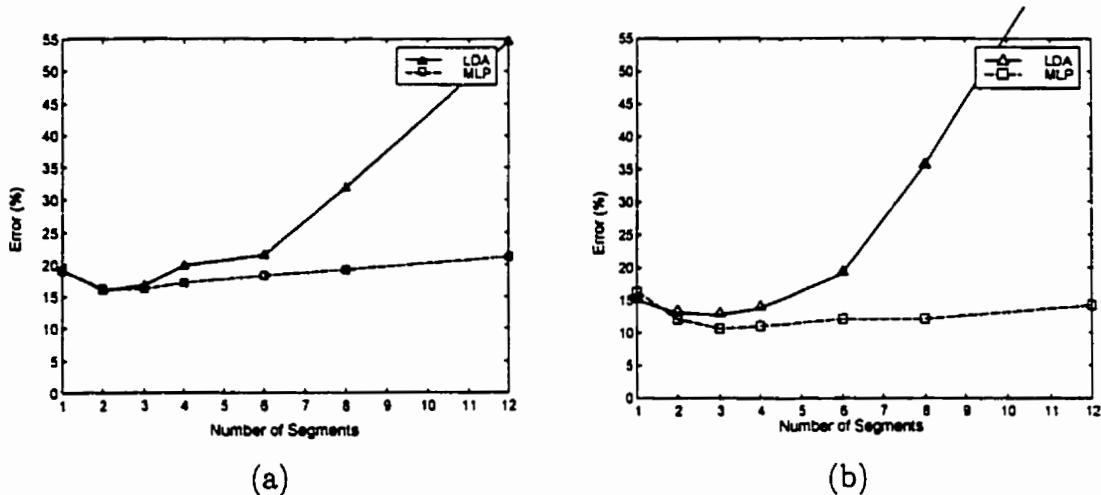


Figure 4.9 -The performance as a function of the number of segments in a $N=240$ record, using LDA and MLP classifiers. Figure (a) depicts the average across all subjects in the one channel dataset, and (b) in the two channel dataset.

The results of these segmentation analyses would seem to indicate that the best generalized segmentation scheme would have 2 or 3 segments of length 120 or 80 ms, respectively¹. At first glance, these results are rather discouraging, implying that the additional temporal structure afforded by a larger number of segments cannot be used effectively. It is important to note however, that when increasing the number of segments within a fixed (240ms) record length, that there are actually two phenomena contributing to degradation of generalization:

1. Increased variance in each feature estimator.
2. The curse of dimensionality.

¹ These results are somewhat different than those reported by Hudgins *et al.* [Hudgins93], in which the optimal segmentation was five 40 ms segments in a 200 ms record when using a MLP classifier. Although these results are not directly comparable (a 240 ms record is being used here), the differences are likely due to the fact that Hudgins' analysis was based upon only six of the 13 subjects in the one channel database.

It is not possible to reduce the effects of feature variance, but one can compensate for the effects of the curse of dimensionality. This is evident in the clearly superior performance of the MLP at higher levels of segmentation, due to the relative immunity of the MLP to the effects of the curse of dimensionality. By applying a scheme of dimensionality reduction, the best segmentation scheme will be determined largely by the effects of estimation variance, and not by the effects of the curse of dimensionality. This will be investigated in the next section “Dimensionality Reduction”.

4.2.3 Dimensionality Reduction

The simplest means of demonstrating the effects of dimensionality reduction is to consider discarding some time domain features of questionable value. Hudgins [Hudgins91] remarked that the MAVS and SSC features appeared to be quite noisy (due to their derivative nature) and scored relatively low in the ANOVA analysis. The waveform segmentation analysis of the one channel data was repeated, omitting the MAVS and SSC features. Figure 4.10 (a) and (b) show the performance with and without these features, using LDA and MLP classifiers respectively.

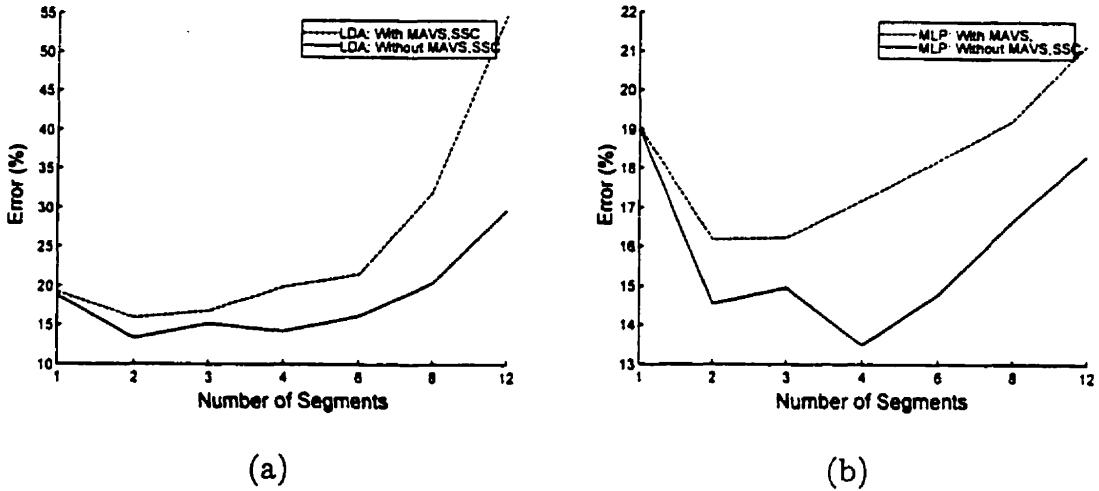


Figure 4.10 – One channel data: the effect of omitting the MAVS and SSC waveform statistics. Figure (a) depicts the error when using a LDA, and (b) when using a MLP.

From these results, three observations are evident:

1. The omission of MAVS and SSC results in improved test set classification, overall.
2. When using the LDA classifier, this improvement is more pronounced in representations using more segments (higher feature set dimensions).
3. The improvements due to discarding these features have a greater effect upon the LDA than the MLP, since the LDA is more susceptible to the curse of dimensionality. This brings the LDA performance closer to that of the MLP.

This is a crude means of dimensionality reduction, but it effectively demonstrates the benefits to be gained. Systematic methods of dimensionality reduction promise to further improve generalization performance. This section will demonstrate that improved classifier generalization is possible through the use of feature selection and feature projection. In addition, the relative merits of these methods will be established.

4.2.3.1 Feature Selection

To demonstrate the process of feature selection, consider first the data from a single subject in the one channel dataset. In *the feature extraction* stage, the full time domain set was extracted from six 40 ms segments², yielding an *original feature set dimension* of $M = 35$. These features may be ranked according to their perceived importance, using a class separability (CS) measure, or a probability of error based measure, either add-on (AO) or knock-out (KO). Table 4.2 portrays the “best” five and “worst” five features for this subject, using each of the three measures of feature importance.

| Rank | Measure of Feature Importance | | |
|------|-------------------------------|---------|---------|
| | CS | AO | KO |
| 1 | WL(3) | MAV(3) | MAV(5) |
| 2 | MAV(4) | WL(7) | WL(6) |
| 3 | MAV(3) | MAV(7) | WL(3) |
| 4 | MAVS(3) | ZC(1) | MAV(4) |
| 5 | MAV(5) | MAVS(3) | WL(5) |
| : | : | : | : |
| 31 | MAVS(2) | WL(2) | MAVS(2) |
| 32 | MAVS(1) | WL(1) | MAV(1) |
| 33 | MAV(1) | MAVS(7) | MAVS(1) |
| 34 | MAVS(7) | MAVS(1) | MAVS(7) |
| 35 | WL(1) | MAV(1) | WL(1) |

Table 4.2 – The “best” and “worst” features for Subject 5, using all time domain features, evaluated upon six 40 ms segments (including an additional segment of mean features, denoted as segment seven). Feature importance has been evaluated using class separability, add-on and knock-out methods.

The feature ranking for this subject has been shown graphically in Figure 4.11, which portrays the rank for each feature separately, across each of six segments (and the seventh, averaged segment).

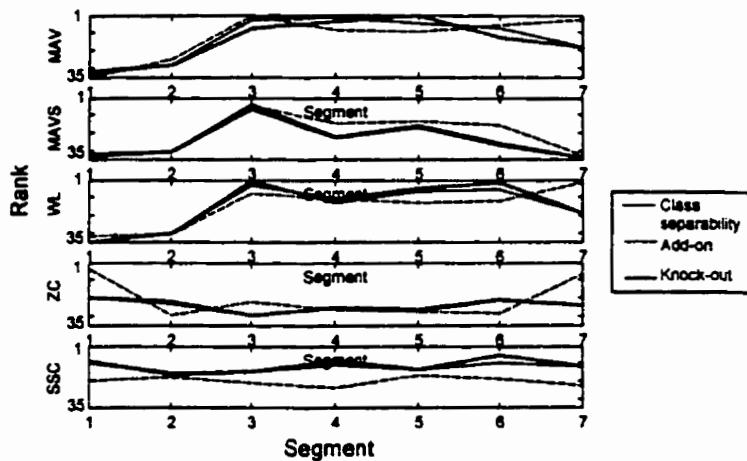


Figure 4.11 – A graphical portrayal of the feature ranking for Subject 5, using all time domain features, evaluated upon six 40 ms segments (including an additional segment of mean features, denoted as segment seven).

From the data in Table 4.2 and Figure 4.11, it is clear that the measures of feature importance yield similar trends when ranking the features. As would be expected, though, there are slight differences due to their objective functions. The effects of these differences upon classification rate will be examined later. The most notable trend is that the MAV, MAVS and WL features seem to provide the *most* and the *least* information; these features tend to be information rich in segments 3-5, and devoid of information in segments 1-2. ZC and SSC tend to provide a moderate contribution throughout all segments; ZC seems to be the least informative feature, overall.

The trends apparent in this dataset are evident throughout the subject roster. Figure 4.12 shows the feature ranking of all subjects, with the mean response superimposed. Class separability has been used as the measure of feature importance³.

² A relatively large number of segments have been chosen here to demonstrate the advantages to be gained by dimensionality reduction.

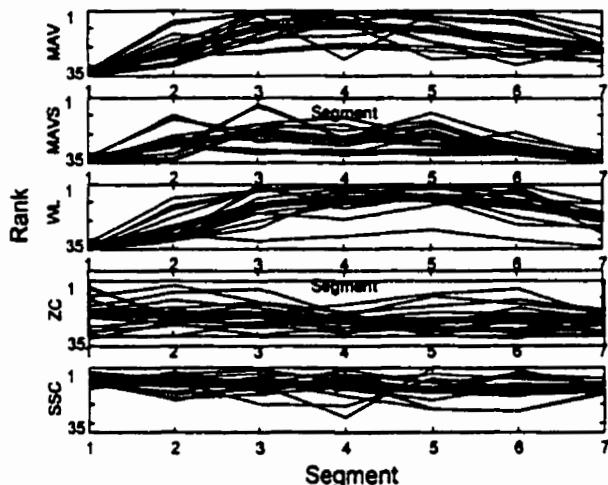


Figure 4.12 – A graphical portrayal of the feature ranking for all subjects, using class separability. The heavy line represents the mean ranking across all subjects.

Clearly, there is an appreciable degree of variability between subjects, but the same trends are evident in the mean ranking as were noted with Subject 5. The implication of the variability is that no one feature should be ignored completely, nor should any segment be neglected altogether. Some generalizations of feature importance may be formed from the ensemble average of the ranking across subjects but, due to a significant degree of inter-subject variability in feature ranking, this ranking should be done on an individual basis.

By ranking these features, we may form lower dimensional feature subsets by choosing only the L most important features. A proper determination of the “optimum” dimensionality of this feature subset must be done using a validation set. Therefore, an examination of the test set classification performance using each feature selection scheme was restricted to the two channel data. Figure 4.13 shows the test set classification error using each feature selection scheme.

³ Add-on and knock-out methods generate responses that exhibit very similar trends.

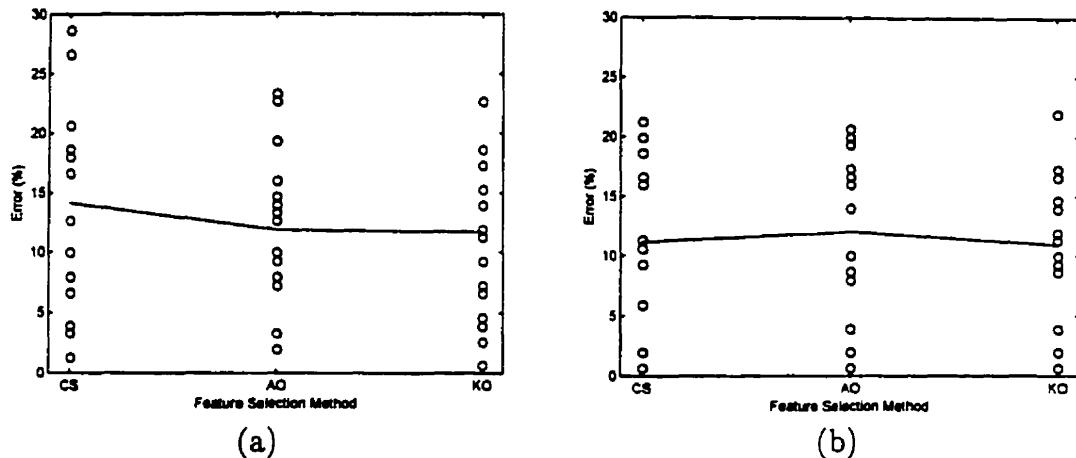


Figure 4.13 – Two channel data: the test set classification performance using subsets of time domain features, determined by CS, AO and KO methods. For each method, the scatterplot of all subjects is superimposed by the mean response across all subjects. Figure (a) shows the response for using a LDA classifier, and (b) using a MLP classifier.

Given the inter-subject variability evident in the scatterplots, there does not appear to be any distinct advantage to any of CS, AO or KO, regardless of the classifier used. Consequently, there is no strong motivation for using AO or KO, which are more computationally intense. Therefore, CS will be used henceforth as the method of feature selection when using TD features, unless otherwise indicated.

4.2.3.2 Optimal Segmentation with Feature Selection

The effects of record length and waveform segmentation were investigated using the full time domain feature set at the beginning of this section. These analyses are now revisited, armed with feature selection methods.

Record Length

Using a class separability distance measure as the basis of feature selection, the record length analysis was repeated. For each subject, the "optimal" feature set dimension was chosen as that which minimized the error upon a validation data set. The test set error was then evaluated at this dimension. The necessity of having a validation set to perform these forms of dimensionality reduction precluded the use of the one channel data due to an insufficient dataset size. Figure 4.14 shows the test set error averaged across all subjects in the two channel dataset, evaluated at each record length. Figure 4.14 (a) depicts the results using the entire feature set, repeated here for convenience. Figure 4.14 (b) shows the results when using CS based feature selection.

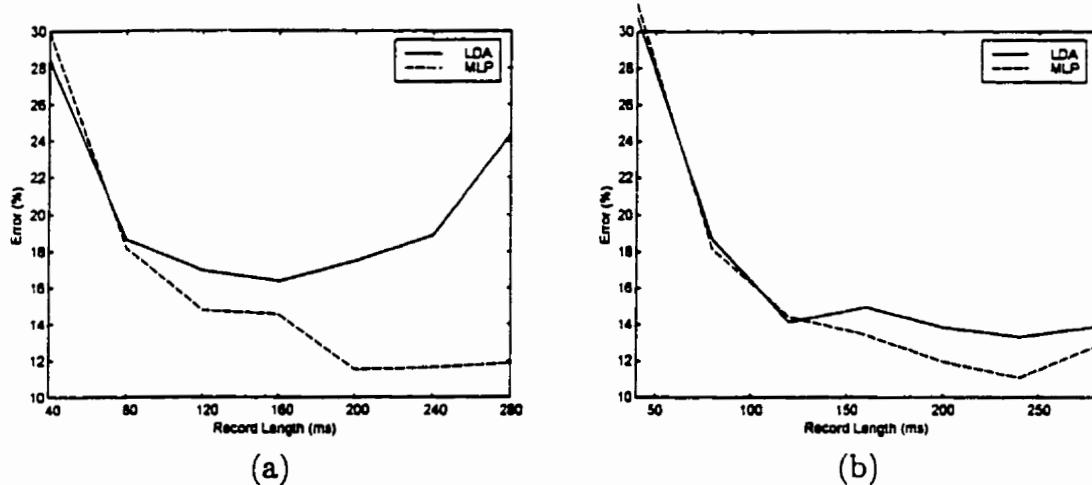


Figure 4.14 – Two channel dataset: the effect of record length upon the test set classification rate, averaged across all subjects. Figure (a) shows the results when using the full feature set, and (b) when using feature selection.

When using a LDA classifier, the deleterious effects of the curse of dimensionality are effectively accommodated by feature selection. When using a MLP, there is little improvement when using feature selection due to the ability of the MLP to accommodate noisy or uninformative features. When using feature selection, a 240 ms record length yields the best test set classification rate, for both classifiers.

Waveform Segmentation

An analysis of the effects of waveform segmentation was performed, complemented by class separability feature selection. As before, the locus of segmentation schemes having a record length of $N = 240$ were considered. Figure 4.15 shows the averaged test set classification error of the two channel dataset, evaluated using each of the segmentation schemes. Figure 4.15 (a) shows the results when using the full feature set, and (b) when using CS based feature selection.

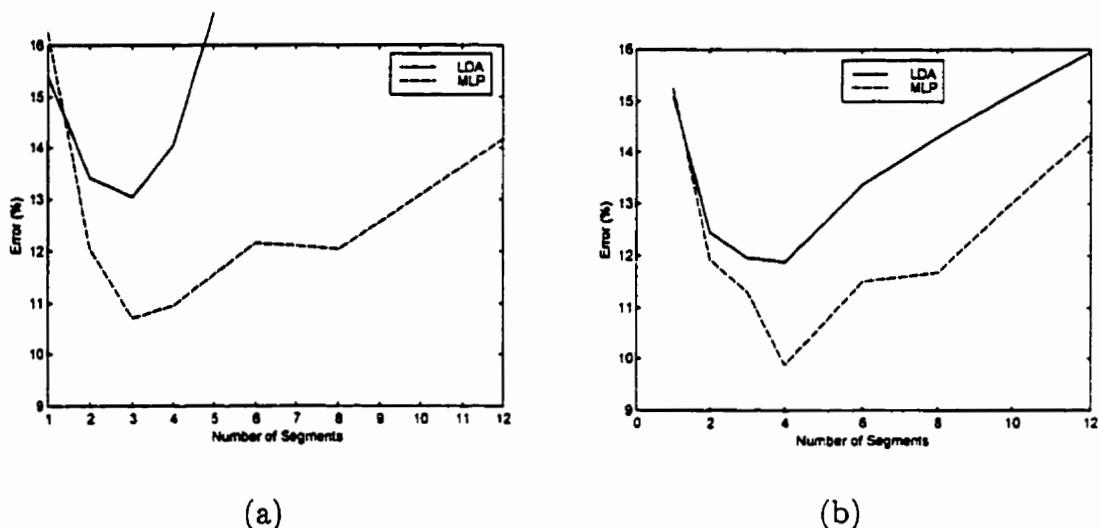


Figure 4.15 – Two channel data: the effects of segmentation upon an $N=240$ record, averaged across all subjects. Figure (a) shows the results when using the full feature set, and (b) when using feature selection.

Feature selection serves to decrease the error rate overall, and to improve the performance with a greater number of segments. This effect is especially dramatic when using the LDA classifier. When using feature selection, four segments of 40 ms seem to yield the best generalization performance for both classifiers.

4.2.3.3 Optimal Segmentation with Feature Projection

It has been shown in the previous section that feature selection can significantly alter the effect that record length and waveform segmentation has upon the generalization performance. PCA-based feature projection will now be examined in the same manner.

Record Length

For each subject the original time domain features were projected onto their principal components; the “optimal” dimension of this projected feature set was chosen as that which minimized the error upon a validation data set. The test set error was then evaluated at this dimension. Figure 4.16 shows the test set error averaged across all subjects in the two channel dataset, evaluated at each record length.

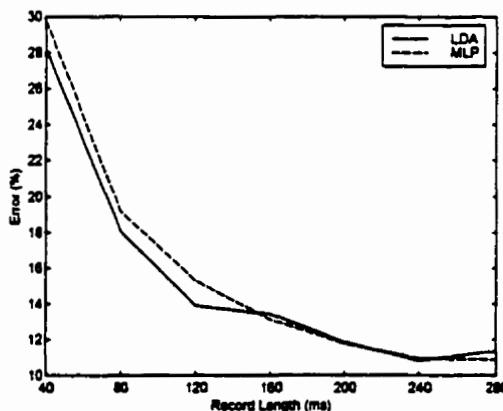


Figure 4.16 – Two channel dataset: the effect of record length upon the test set classification rate when using PCA feature projection

The result is similar to that when using feature selection; the additional information in longer records is assimilated without undue dimensionality. Again, a 240 ms record yields the best classification error using both classifiers.

It has been shown that with an appropriate form of dimensionality reduction (either feature selection or feature projection), the latter segments do indeed contribute useful information. Therefore, subsequent time domain feature extraction will use a 240 ms record, with the expectation that some form of feature reduction will be used.

Waveform Segmentation

With the time domain feature set complemented by PCA feature projection, an analysis of the effects of waveform segmentation was performed. As before, the locus of segmentation schemes having a record length of $N = 240$ were considered. Figure 4.17 shows the averaged test set classification error of the two channel dataset, evaluated using each of the segmentation schemes.

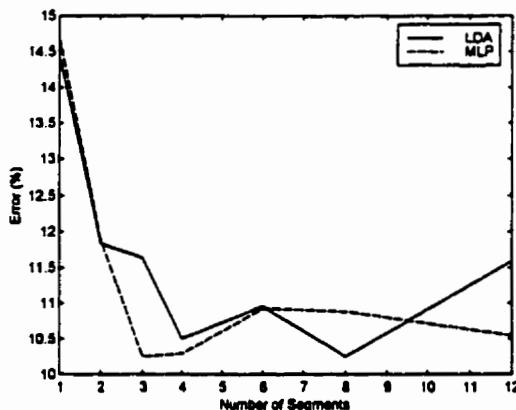


Figure 4.17 – Two channel data: the effects of segmentation upon an $N=240$ record, averaged across all subjects, when using feature projection.

The results are similar to those produced when using feature selection: the overall performance is improved over the full feature set, and a greater degree of segmentation is supported. Both classifiers perform well when using four segments, although the LDA does slightly better when using eight. Some

distinctions exist, however. The most notable difference is that even with very large numbers of segments (greater than 6), the error rate remains quite low. This is attributable to the PCA's ability to synthesize information that may be dispersed, as is the situation with a highly segmented waveform.

4.2.4 Summary

When using time domain features as a basis of signal representation for classification, it has been shown that proper specification of record length and waveform segmentation is essential. In the process of this investigation, it has become evident that generalization performance suffers due to the curse of dimensionality. Dimensionality reduction methods promise to improve classifier generalization and concurrently, alter the performance with respect to record length and waveform segmentation. Dimensionality reduction subtends better performance by allowing more information to be conveyed in a lower dimensional space:

- Longer record lengths can be accommodated, thereby allowing the information in the latter portion of the record to be used.
- A greater degree of waveform segmentation is possible, allowing the representation to better capture the temporal structure of the data.
- Irrelevant and distracting features are ignored or de-emphasized.

At this point it would be useful to provide a direct comparison of the classification performance when using the full feature set, when using feature selection, and when using feature projection.

Figure 4.18 shows the effects of segmentation when using each form of dimensionality reduction and when using the full feature set.

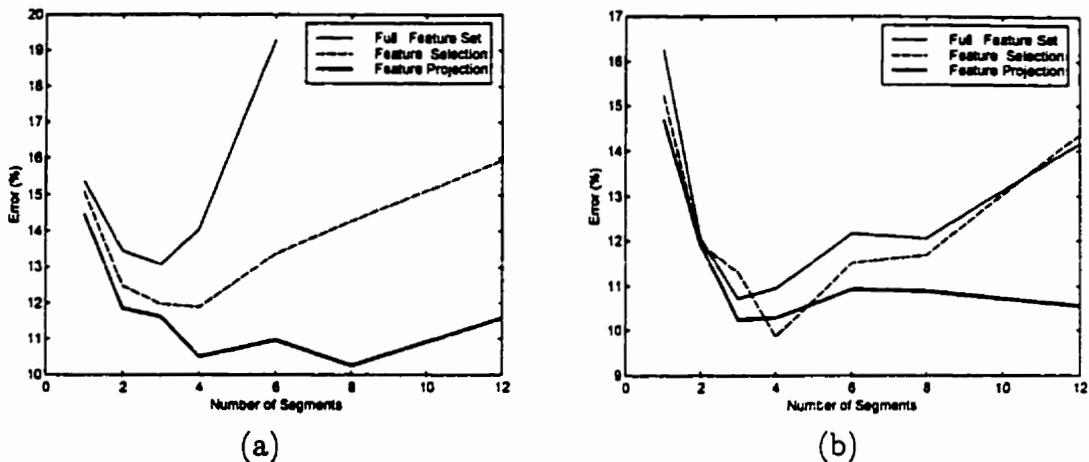


Figure 4.18 – Two channel data: a comparison of dimensionality reduction schemes with respect to waveform segmentation, using a $N=240$ ms record. Figure (a) depicts the performance when using a LDA classifier, and (b) when using a MLP classifier.

The effects of dimensionality reduction are obvious when using a LDA; in Figure 4.18 (a) feature selection is clearly superior to no dimensionality reduction, and feature projection outperforms feature selection. The benefits are less pronounced but still present when using a MLP, as shown in Figure 4.18 (b). Although feature selection provides the lowest overall error (at four segments), feature projection maintains good performance over a wide range of segmentation schemes. A scheme employing four segments provides the best (or near-best) generalization for both methods of dimensionality reduction.

To further focus upon the relative performance of these methods, the best segmentation scheme was used for the full feature set (three segments), feature selection (four segments) and feature projection (four segments). Figure 4.19 shows the relative performance of each scheme.

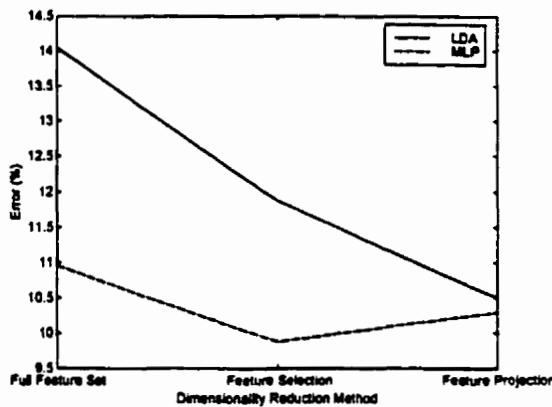


Figure 4.19 – Two channel data: the relative performance of using optimal segmentation scheme. This has been done when using the full feature set (three segments), feature selection (four segments) and feature projection (four segments).

The LDA classifier clearly benefits most from feature projection, while feature selection appears to have a slight advantage when using the MLP classifier. It is interesting that the performance of the LDA can approach that of the MLP when using feature projection. The reason for exceptional performance when using the PCA / LDA combination is that PCA removes the linear dependencies and “noisy” features from the data, which may distract the LDA classifier.

4.3 STFT Based Features

The MES classification problem is quite distinct from those involving many physical signals; the nature of this distinction will become apparent when transforming these signals into a time-frequency representation. Most signals of physical origin have a significant degree of structure in the time domain, the frequency domain, or both. Such structure allows the signal energy to be concentrated into a reasonably defined subset of time-frequency coefficients, and therefore feature selection performs well as a means of determining a reduced feature set for classification. Some examples of signals exhibiting significant time-frequency structure include speech (phoneme utterances), short-duration underwater acoustic signals, radar, and most detection problems involving signals contaminated by additive noise.

The transient myoelectric signal however, has a large degree of structural variance that is quite different than additive noise, amplitude or frequency modulation, filtering, or other modifying influences that may contribute to within-class variance. The structural variance is due to the fact that the surface MES arises from a multitude of semi-random sources (motor units). Each source, in turn, is subject to further modification during volume conduction prior to measurement. Therefore, although patterns within a given class possess visually perceptible similarities, the signal energy is highly dispersed in time and in frequency. This time-frequency dispersion presents a difficult signal representation problem, rendering TFR *feature selection* methods relatively ineffective. *Feature projection* and averaging methods are more successful because they synthesize this dispersed information.

This section provides an introduction to TFR based feature sets, investigating the capabilities of the STFT as a basis for transient MES classification. Many of the concepts and methods common to subsequent sections dealing with TFRs will be introduced.

4.3.1 Related Work

The STFT is a natural extension of the discrete Fourier transform, and therefore is well understood, is easily interpreted, and has highly efficient fast algorithms. For these reasons, the STFT has been of great interest in many diverse fields in which the signals of interest possess time-frequency structure. The STFT is extremely important as a data visualization tool, due to its intuitive representation and its computational efficiency. In signal analysis, it provides diagnostic information not easily measured or interpreted solely in the time or frequency domain.

Although the origins of the STFT stem from the early work of Gabor [Gabor1946], its use as a basis for pattern recognition is a relatively recent development. The research that has drawn most heavily upon the STFT is that of speech recognition; indeed, the spectrogram has been casually referred to as a “voiceprint.” Phoneme sounds and continuous speech are well concentrated in the time-frequency domain, and as a result, the STFT has proven an effective signal representation for classification [Intrator91][Waibel89]. Other “large” bodies of research in which the STFT has become an important basis for pattern recognition include underwater acoustic signals (sonar) [Beck94], underground acoustic signals (oil exploration, seismology), and computer-assisted cardiography, using both the electrocardiogram [Dickhaus94] and the phonocardiogram [Durand90].

The STFT has been used as a basis for analyzing the transient MES. Perhaps the first work of this type was that of Hannaford *et al.* [Hannaford86]. The STFT was used to track temporal variations of the MES spectrum accompanying rapid movements of the head and wrist. Although no attempt was made to perform pattern recognition, repeatable structure was identified in the MES spectrogram. Kelly *et al.* [Kelly90] proposed a scheme of classifying the stationary MES from the upper limb using a periodogram and a MLP classifier. The accuracy of this method was limited by the lack of structure in the stationary MES, and the absence of any form of dimensionality reduction strategy. Gallant *et al.* [Gallant93] proposed a method of classifying the transient MES from the upper limb based upon the STFT, coupled with a form of projection pursuit dimensionality reduction. The results were promising, but difficult to interpret in the absence of a direct comparison with other methods. Farry *et al.* [Farry96] have investigated using the STFT to classify the transient MES accompanying rapid motion during grasping and thumb motions. They reported good results on a very limited amount of data (two subjects), and demonstrated the feasibility of real-time implementation using a dedicated DSP processor.

The purpose of this section is to generalize and provide greater insight into classification using STFT features. The effects of STFT parameters and dimensionality reduction strategies will be examined in detail. The choice of each transform-related parameter and the most effective means of dimensionality reduction will be based upon empirical evidence: the effect upon classification performance of the ensemble of two channel transient MES data. It will be shown that for MES classification, the effects of dimensionality reduction are much more profound than any transform-related parameter.

4.3.2 STFT Feature Extraction Parameters

The use of the STFT for any application requires selection of a number of parameters relevant to the feature extraction process. When constructing a feature set for classification, some important *transform-related* parameters are:

1. **STFT window size.** The time-frequency energy signatures of the signals of interest determine the best tradeoff between time resolution and frequency resolution.
2. **STFT window overlap.** It is possible that useful information may be obtained by an overdetermined time-frequency tiling.
3. **STFT window type.** This will affect the manner in which energy is localized in each time-frequency cell.

An exhaustive investigation of STFT parameters would require every possible combination of window size, window overlap and window type; this is not feasible due to the required computational burden. Instead, a sub-optimal search will be performed sequentially:

1. **Determine the best window size.** This will assume a certain window type and overlap. The window size will be selected as that yielding the lowest test set error, assuming that the effect of window size under these conditions is a valid generalization to other window types and overlaps.
2. **Determine the best window overlap,** using the window size selected in Step 1 and an assumed window type. The assumption here is that the classification performance with respect to overlap generalizes reasonably well to other window types.
3. **Determine the best window type,** having selected the best window size and overlap.

This sequential parameter selection has been specified in the order of the anticipated significance: window size is expected to have the most significant effect, and window type the least. Transform parameter selection will likely depend on whether feature selection or feature projection is used as a dimensionality reduction technique. Therefore, a separate set of STFT parameters must be determined for selection based and projection based STFT feature sets. During parameter optimization, class separability (using Euclidean distance) and PCA were chosen to represent feature selection and feature projection, respectively.

For each subject in the two channel dataset, the “optimal” feature set dimension was estimated as that yielding the lowest validation set classification error. The generalization performance was evaluated as the test set classification error at the chosen feature dimension. This analysis has been done using both a LDA and a MLP classifier, to accommodate for differences accountable to the capabilities of each classifier.

4.3.2.1 STFT Window Size

Using a window of length W samples and a sampling rate f_s , one is presented with the following tradeoff in time resolution (Δt) vs. frequency resolution (Δf):

$$\Delta t = \frac{W}{f_s}, \quad \Delta f = \frac{f_s}{W}. \quad (4.6)$$

Consider for now window sizes that are even divisors of the record length, $N=256$.

With no overlap between windows the number of time windows in the TFR is $\frac{N}{W}$, and the number of frequency bins is W . Therefore the dimension of the TFR, M , is equal to the dimension of the original signal space, N . This TFR however, includes the double-sided spectrum. Only one half of the spectrum is needed for pattern recognition, so the effective dimensionality of the critically sampled STFT is

$$M = \underbrace{\frac{N}{W}}_{\text{number of time cells}} \times \left\{ \underbrace{\frac{W}{2}}_{\substack{\text{one-half} \\ \text{of spectrum}}} + \underbrace{\frac{1}{\text{DC component}}}_{\substack{\text{number of frequency cells}}} \right\}. \quad (4.7)$$

The effective dimensionality of the STFT is therefore slightly greater than $\frac{N}{2}$. The table below delineates the possible time-frequency tilings of a critically sampled transform. Included is the dimensionality of the double-sided and single-sided TFRs.

| Δt (ms) ¹ | Δf (Hz) | Double-sided STFT (time windows \times frequency bins= M) | Single-sided STFT (time windows \times frequency bins= M) |
|---------------------------------|-----------------|---|---|
| 4 | 250.0 | $64 \times 4 = 256$ | $64 \times 3 = 192$ |
| 8 | 125.0 | $32 \times 8 = 256$ | $32 \times 5 = 160$ |
| 16 | 62.50 | $16 \times 16 = 256$ | $16 \times 9 = 144$ |
| 32 | 31.25 | $8 \times 32 = 256$ | $8 \times 17 = 136$ |
| 64 | 15.63 | $4 \times 64 = 256$ | $4 \times 33 = 132$ |
| 128 | 7.81 | $2 \times 128 = 256$ | $2 \times 65 = 130$ |
| 256 | 3.91 | $1 \times 256 = 256$ | $1 \times 129 = 129$ |

¹ Since sampling has been performed at $f_s = 1000\text{Hz}$, time measures may be referred to synonymously in terms of samples or milliseconds. Heretofore, any reference to time will carry implicit units of samples (or milliseconds).

The best combination of Δt and Δf for classification of the MES is not immediately obvious. Certainly, selecting a window of $N=256$ specifies a simple periodogram, discarding all time structure. Conversely, when using a window size of 4, the frequency resolution is very poor, and it is unlikely that information localized in frequency can be discriminated. These heuristics however, are meaningless without evidence. The transient MES possesses a complex time-frequency characteristic and therefore, the only means of determining the best time-frequency resolution tradeoff (with respect to classification performance) is by empirical observation.

Figure 4.20 shows the validation set and test set classification error for the range of window sizes. The response represents the ensemble average across all subject data. The assumed parameters here are a Hamming window with no overlap; this is a common default in nonstationary spectral analysis.

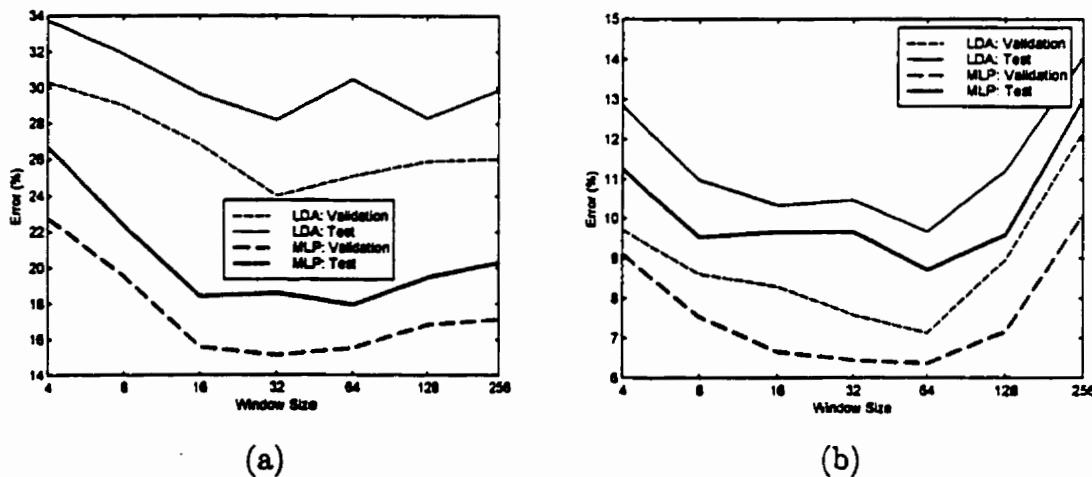


Figure 4.20 – The effect of STFT window size upon classification error; (a) feature selection using class separability, (b) feature projection using PCA.

On the left, the response using feature selection is shown. Although somewhat ambiguous, the best performance seems to result when using a window length of 32. On the right, the response using feature projection differs somewhat: a clear

advantage is evident when using a window size of 64. This is true for both validation and test sets, and for both LDA and MLP classifiers. It is important to note that the classification error is *much* lower when using feature projection as a means of dimensionality reduction. This will be a fundamental observation throughout the investigation of all time-frequency based feature sets. It is also notable that the MLP outperforms the LDA when using feature selection, due to the ability of the MLP to handle higher feature dimensions than the LDA². It is also worthy of note that the validation set consistently outperforms the test set: this is because the estimate of feature dimension is derived from the validation set.

4.3.2.2 STFT Window Overlap

Having selected the window length for feature selection and feature projection methods, the focus now shifts to determining the best window overlap. Overlapping STFT windows subtend TFRs of higher dimension than non-overlapped TFRs; some of the information in an overlapped TFR is redundant, and some of it is novel. If the number of samples by which adjacent time windows overlap is O , the total number of time windows is

$$\text{time windows} = \frac{N-O}{W-O} \quad (4.8)$$

which yields a TFR of dimension

$$M = \left(\frac{N-O}{W-O} \right) \times \left(\frac{W}{2} + 1 \right). \quad (4.9)$$

²The chosen feature dimension when using feature selection is much greater than when using feature projection. This is because the information important for discriminating transient MES patterns is highly dispersed in the time-frequency domain.

The dimensionality of the TFR as a function of overlap is given in the table below (for a signal length of $N = 256$ and window sizes $W = 32$ and $W = 64$):

| Overlap (O) | TFR Dimension, M ($W = 32$) | TFR Dimension, M ($W = 64$) |
|--------------------|------------------------------------|------------------------------------|
| 0 | 136 | 132 |
| $\frac{W}{4}$ | 175 | 165 |
| $\frac{W}{2}$ | 255 | 231 |
| $\frac{3W}{4}$ | 493 | 429 |

Again using a Hamming window, the validation set and test set classification errors were determined for the range of window overlap $O = \left[0, \frac{W}{4}, \frac{W}{2}, \frac{3W}{4}\right]$. The response, averaged across all subjects, is shown in Figure 4.21.

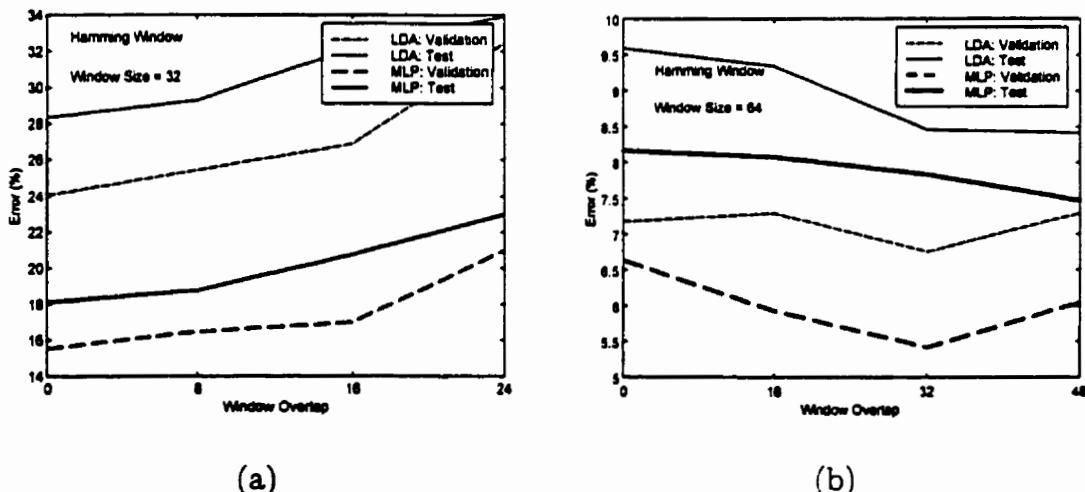


Figure 4.21 – The effect of STFT window overlap upon classification error; (a) feature selection using class separability, (b) feature projection using PCA.

When using feature selection (and a window size of $W=32$), it appears that the best performance results when using no overlap. When using feature projection (and a window size of $W=64$), an overlap of $O = \frac{W}{2}$ (32 samples) seems to yield the best classification error, on average. Feature projection effectively synthesizes the added information embedded in higher dimensional TFRs, whereas feature selection does not. Increasing the overlap from $O = \frac{W}{2}$ to $O = \frac{3W}{4}$ does not seem to subtend a significant amount of information important for classification. Indeed, the TFR is of sufficiently high dimension (429×2 channels = 958) that it tends to overwhelm even the PCA. As well, the determination of the principal components (in the training phase) of such a large feature set requires substantial computational effort.

4.3.2.3 STFT Window Type

The final STFT related parameter to be “tuned” to transient MES analysis is that of window type. The necessity of windowing the signal when computing a DFT imposes two deleterious effects: *reduced spectral resolution* and *spectral leakage* [Oppenheim89]. The resolution is primarily influenced by the width of the mainlobe of the window’s Fourier transform. This describes the frequency localization characteristics of the resulting transform. The leakage depends upon the amplitude of the window’s sidelobes, relative to the amplitude of the mainlobe. This introduces a bias into the transform.

Many window functions (or *tapers*) have been introduced to control the effects of bias in spectral estimation. A rectangular (boxcar) window has minimal mainlobe width (maximum resolution) and maximal bias (severe sidelobe distortion). All

other window types necessarily sacrifice resolution to reduce the bias in the transform. A representative sample of window types are shown in Figure 4.22; the Bartlett (triangular), Blackman, Hanning, and Kaiser windows (for details see [Marple87]). The left column shows the time envelope, and the right column shows the logarithmic frequency response. The windows are presented in order of increasing mainlobe width, and consequently, decreasing sidelobe distortion.

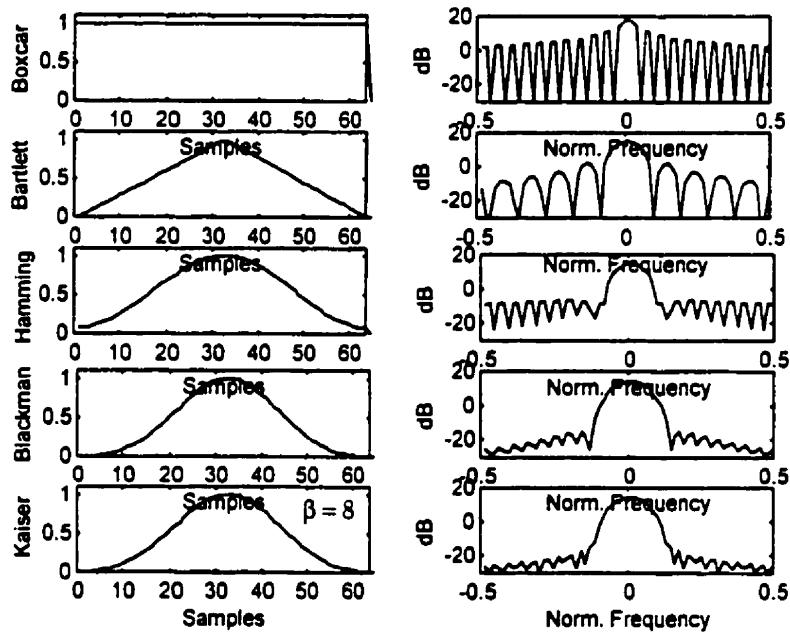


Figure 4.22 — The temporal envelope and frequency response of some common smoothing windows.

The Kaiser window has a parameter (β) that is used to adjust the bias versus resolution tradeoff. A value of $\beta = 8$ has been chosen to subtend a window type with low bias.

In addition to classical taper windows, a more recently developed nonparametric spectral estimator was investigated. Thomson's Multiple Window Method (MWM) promises to improve the bias-variance tradeoff of the periodogram without imposing a misleading parametric model [Thomson85]. Instead of

applying one *ad-hoc* window to the signal, a *set* of windows is applied that are (i) mutually orthogonal and (ii) optimally concentrated in frequency. These windows – called *discrete prolate spheroidal sequences* (DPSSs) or *Slepian sequences* – are the eigenvectors of the operation corresponding to a cascaded band-limiting to $[-B, B]$ and time-limiting to $[0, N - 1]$ [Slepian61].

For a single segment of length L , the MWM transform is

$$P_{MWM}(f) = \frac{1}{2K} \sum_{k=0}^{K-1} \frac{1}{\lambda_k} |x_k(f)|^2, \quad (4.10)$$

where $K = 2LB$ and $x_k(f)$ are the eigenspectra

$$x_k(f) = \sum_{\ell=0}^{L-1} x[\ell] v_k[\ell] e^{-j2\pi f\ell}. \quad (4.11)$$

The $v_k[\ell], k = 1, \dots, K$ are the DPSS windows and λ_k are the corresponding eigenvalues. Only the largest K DPSS windows are used in the estimate: the parameter K controls the tradeoff between bias and variance. Increasing values of K (and therefore, B) subtend lower variance at the expense of greater bias. Typical values of K range from four to ten. The MTM can be easily extended to yield a time-frequency representation (the *short-time Thomson transform* – STTT) just as the periodogram is translated to yield the STFT.

Farry has applied the STTT to transient MES signals from the hand muscles accompanying grasping motions [Farry96]. She has shown that the STTT does indeed possess lower variance than the STFT applied to these signals. Classification of motion type gave a slight advantage to the STTT over a Hamming windowed STFT for the single subject study. A simple bin averaging scheme was used to perform dimensionality reduction.

The STTT was applied to the ensemble of two channel data of this work, and feature selection and feature projection dimensionality reduction were applied to the STTT TFRs. There was no marked difference in test set classification error amongst values of K through the range [4-10], but a value of 8 seemed to confer a slight advantage across the ensemble average.

Each of the STFT taper windows previously mentioned and the STTT were evaluated in the context of transient MES classification. In each case, the STFT window size and overlap were set to their “optimum” values, as previously determined. Figure 4.23 shows the validation set and test set classification error for each window type.

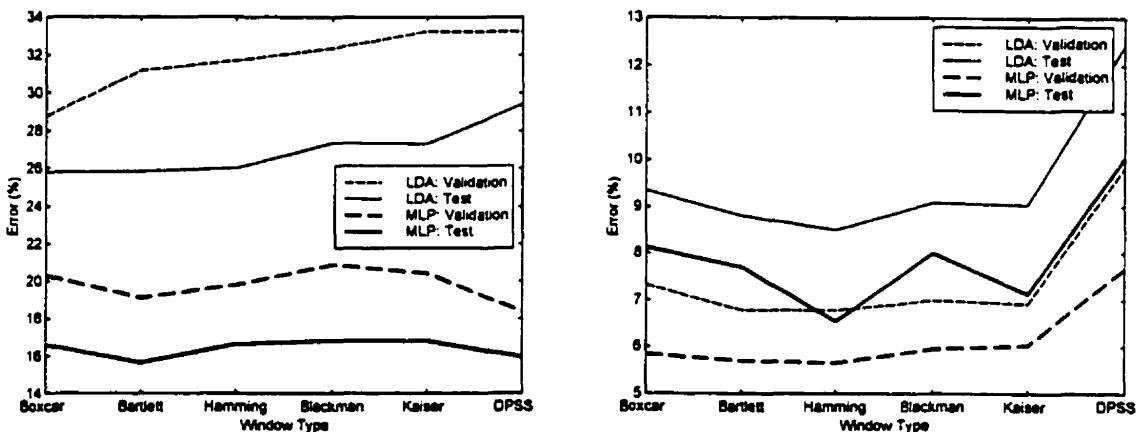


Figure 4.23 – The effect of window type upon classification error; (a) feature selection by class separability, (b) feature projection using PCA.

The plots have been presented such that STFT taper windows possess decreasing resolution (and lower bias) when progressing from left to right on the horizontal axis. The STTT (DPSS window) result has been shown to the extreme right. As expected, there is not a dramatic effect attributable to window type. When using feature selection, the MLP seems to slightly favor the Bartlett and DPSS windows.

Interestingly, the LDA does progressively worse with windows of lower resolution (and lesser bias), exhibiting its best performance with a boxcar window. When using feature projection, the best performance comes from those windows that offer a balance of resolution and bias. The Hamming window seems to hold a slight advantage, for both the LDA and the MLP. The STTT does relatively poorly in combination with feature projection; the reason for this has not been determined.

The following is a summary of the analysis of transform based parameters when using the STFT for transient MES classification.

1. Transform based parameters must be determined in the context of the means of dimensionality reduction. Analyses were performed here using a example of feature selection (class separability) and feature projection (PCA). Results were interpreted via the average response across the ensemble of all subject data.
2. The best window size is $W=32$ samples when using feature selection, and $W=64$ samples when using feature projection.
3. The best overlap is none at all when using feature selection and $\frac{w}{2}$ (32 samples) when using feature projection.
4. The best window types when using feature selection are the boxcar (LDA classifier) and the Bartlett (MLP classifier). When using feature projection, the Hamming window offers slightly better performance for both classifiers.
5. The combined effects of window size, overlap and type are much less significant than the effect of dimensionality reduction. Feature projection dramatically outperforms feature selection.

The next section examines dimensionality reduction strategies in greater detail.

4.3.3 Dimensionality Reduction Methods

In the previous section, feature selection and feature projection were used to contrast the performance of transform based STFT parameters. This section will provide an investigation of dimensionality reduction in greater detail.

4.3.3.1 Feature Selection

Three feature selection methods have been previously described; class separability (CS), add-on (AO) and knock-out (KO). In this section, they are compared with respect to their ability to reduce the STFT for the purpose of classifying the transient MES. The best STFT parameters were determined in the previous section in the context of CS based feature selection. These were:

Window size: 32 ms (samples)
Window overlap: 0
Window Type: boxcar (LDA), Bartlett (MLP)

It has been assumed that these parameters are a fair generalization for AO and KO as well. Using these “tuned” STFT parameters, the performance of CS, AO and KO feature selection was compared. Figure 4.24 shows the test set error rate (averaged across all subjects), evaluated at all possible feature set dimensions³.

³ Actually, only the feature set dimensions are shown that contribute information. When using the LDA, the error rate continues to grow with increasing dimension, so dimensions higher than 64 are not shown. When using a MLP, the error rate tends toward an asymptote at a dimensionality of 120.

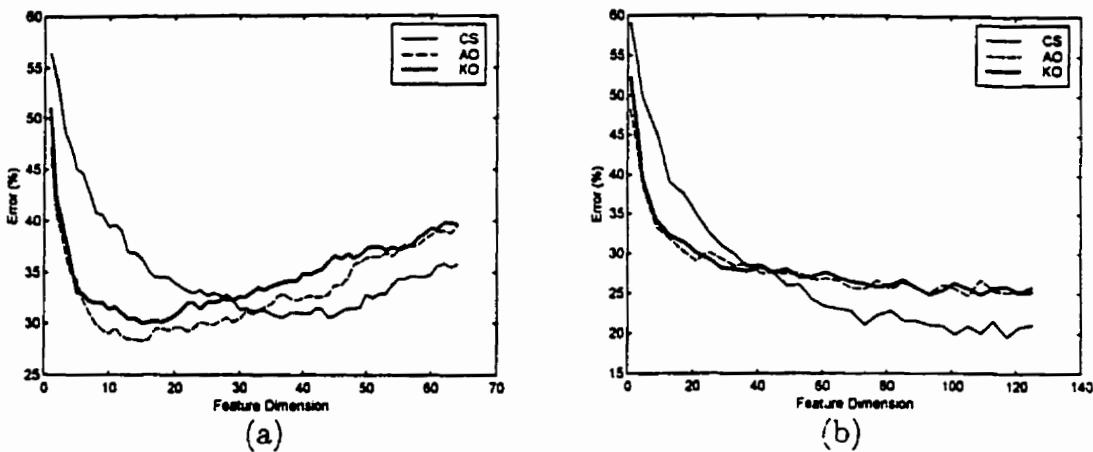


Figure 4.24 – The performance of three feature selection schemes: class separability, add-on and knock-out. Figure (a) depicts the performance when using a LDA classifier; (b) depicts the performance when using a MLP classifier.

From the response in Figure 4.24(a), it appears that, when using a LDA classifier, the add-on and knock-out methods perform slightly better than does class separability. None of the methods, however, perform well in the absolute sense, and all are prone to degradation as feature dimension grows large. The classification performance is somewhat better when using a MLP classifier, as is evident in Figure 4.24(b). Here, the effects of the curse of dimensionality are less severe – higher feature dimensions can be accommodated. As error rates settle asymptotically at a dimension of about 120, class separability seems to outperform the add-on and knock-out methods.

A more meaningful measure of performance is the test set error, evaluated at the “optimal” feature set dimension, as estimated from the validation set error. Figure 4.25 shows the test set error for each of the feature selection methods. The scatterplot of the responses from all subjects have been included, superimposed by the mean of the overall behavior.

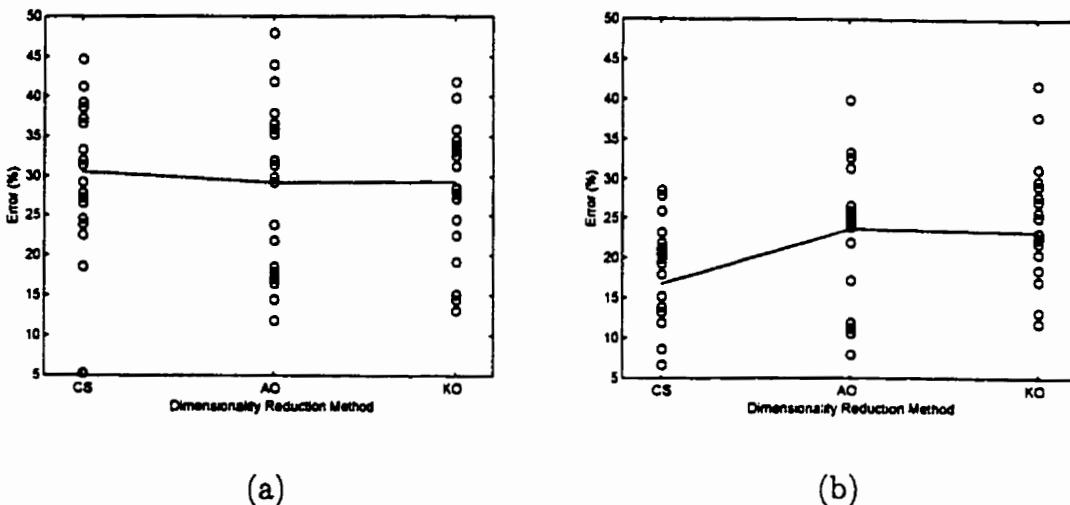


Figure 4.25 ... The test set error rate subtended by each feature selection method, evaluated at the “optimal” feature set dimension. The response is shown (a) when using a LDA classifier, and (b) when using a MLP classifier.

From Figure 4.25(a), it is apparent that little difference exists amongst the methods when using a LDA classifier. In Figure 4.25(b), class separability seems to enjoy a slight advantage when using a MLP classifier.

Since no significant improvement is gained by using AO/KO methods over CS, the simpler CS method will be the feature selection method of choice here, unless otherwise indicated.

4.3.3.2 Feature Projection

The best STFT parameters were determined in the previous section in the context of PCA based feature projection. These were:

Window size: 64 ms (samples)
 Window overlap: 32 ms (samples)
 Window Type: Hamming (LDA), Hamming (MLP)

Using these “tuned” STFT parameters, the performance of PCA based feature projection is shown in Figure 4.26 as a function of feature set dimension.

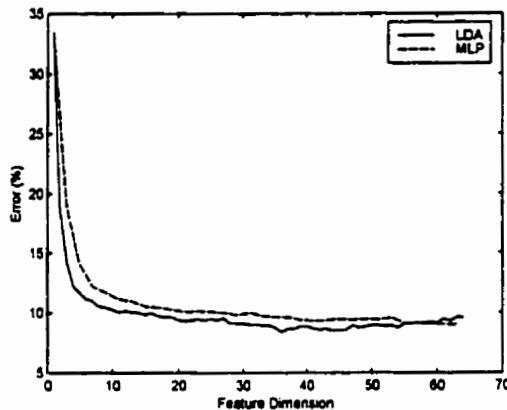


Figure 4.26 – The performance of PCA based feature projection with respect to feature set dimension. The results, averaged across all subjects, are shown for both LDA and a MLP classifiers.

Clearly, very good classification performance is possible with very few features, as compared to feature selection. Overall, PCA feature projection dramatically outperforms feature selection, subtending an averaged classification error that is roughly half that of the best-case feature selection scenario.

The question is: why is PCA, which is supposedly more adept for signal reconstruction than classification, such an effective dimensionality reduction strategy here. The reasons for the superiority of PCA to feature selection when classifying the transient MES are twofold.

1. *The projected features are mutually uncorrelated*⁴. By projecting the data onto the orthonormal axes of maximum variance (the eigenvectors of the covariance matrix), the covariance structure is removed [Fukunaga90]. If

⁴ Indeed, if the original features are Gaussian, the projected features are mutually *independent*.

there are significant linear dependencies in the original feature space, then it may be possible to discard most of the lesser principal components with little loss of information. In the situation where information is liberally dispersed amongst the original feature set, a PCA will consolidate this information much more effectively than feature selection.

2. *The method is unsupervised.* Although it may seem counterintuitive, the knowledge of class membership may actually deteriorate the efficacy of a dimensionality reduction technique. This is because embedding class membership information into the method will bias the representation to the training data in the same manner that a classifier may be biased, hampering the generalization performance. Each of the feature selection methods discussed here rely upon class membership in their feature evaluation criteria. PCA, on the other hand, uses no prior knowledge of class membership, and does not experience bias toward the training set. If the variance in the data can be explained by the signal (rather than the noise), then the leading principal axes will tend to pick projections with good separations.

When using the STFT of the transient MES, the advantages of PCA over feature selection are dramatic, much more so than when using time domain features. This implies that the STFT contains a significant amount of linear dependency amongst the coefficients. As well, the loose structure of the transient MES subtends a substantial degree of within-class dispersion in the time-frequency domain. PCA appears to effectively accommodate these effects. Conversely, feature selection requires so many features to provide adequate discrimination that the resulting dimensionality subtends poor generalization. The improvement that PCA offers to

TD features is not as pronounced as that given to the TFR sets, as the original dimensionality is relatively low.

Of particular interest as well is that, when using the STFT, the LDA classifier actually generalizes as well as the MLP classifier when using PCA. This is despite the fact that the MLP enjoys the advantage over the LDA of being capable of prescribing nonlinear class boundaries. To explain the LDA's performance, consider an arbitrary low-dimensional signal representation in which the class boundaries are indeed nonlinear. In this situation, a MLP will most certainly outperform a LDA. Now consider partitioning the feature space (in time, frequency, or some other domain) such that a larger feature set is formed. As the feature set dimensionality grows, the degree of nonlinearity between class boundaries must diminish. In the high dimensional feature space of the STFT, it is unlikely that highly nonlinear bounds exist between the classes. If a significant degree of linear dependency exists as well, a PCA will project the STFT coefficients onto a relatively low dimensional space, while preserving the linearity that exists between classes in the higher dimensional space. The fact that the PCA-projected STFT features have reasonably linear class boundaries and that they have relatively low dimension diminishes the advantage that a MLP may have over a LDA.

A MLP that is appropriately trained and that has an appropriate number of hidden layer nodes will always match, if not exceed, the performance of a LDA. Due to the need to automate MLP training over a large number of iterations however, the hidden layer size was fixed at eight and the stopping criterion was a fixed number of iterations (200). These generalizations were determined by empirical analysis of the validation set data. For a given subject however, the size

of the MLP may be suboptimal, or the network may be slightly overtrained or undertrained. Both of these factors will inhibit the generalization performance of the MLP. The LDA does not require heuristic specification of its architecture or training algorithm, yet it consistently performs very well.

The advantage that PCA does offer to MLP classifiers is with respect to training time. A speedup of the backpropagation algorithm may result from the application of PCA since the Hessian matrix of the cost function is more diagonalized than usual. This generates an appropriate scaling of the learning rate along each weight axis independently.

It will be shown in the following sections that these observations which relate the performance of PCA versus CS (and also LDA versus MLP) will recur in the investigation of WT and WPT based feature sets.

4.3.3.3 Bin Averaging

Bin averaging is an heuristic approach to dimensionality reduction; time-frequency cells are combined (averaged) according to some *a priori* knowledge (or intuition) of the energy distribution of the signals of interest. Many averaging schemes exist, due to the multitude of ways that signal energy may be distributed in the time-frequency plane.

Although bin averaging is obviously a trivial subset of systematic methods of feature projection, two useful strategies have been treated individually here for demonstrative purposes. A generalized scheme commonly referred to as *Melscale* (or *Mel-warping*) models the acoustic response of the human cochlea [Waibel89].

The frequency response is divided into subbands, which provide a logarithmic coverage of the frequency domain in the same manner as a wavelet basis. In the simplest implementation, the lowest STFT band is kept intact. While increasing in frequency, the next two subbands are averaged, the next four subbands are averaged, and so on ... until the highest subbands are accommodated.

Farry *et al.* [Farry96] have used a more empirical approach to reduction of the transient MES STFT. They noted that the lower portion of the transient MES spectrum was highly variable, and therefore, was of little use in pattern recognition. Similarly, the high end of the frequency spectrum was observed to offer little information. Consequently, only the middle portion of the spectrum was retained – that residing in the 75-250 Hz passband. Farry’s STFT used six 40ms windows (padded to 64 samples) and a sampling rate of 1000Hz. This passband corresponds to 12 coefficients, which were further reduced to four by averaging adjacent triplets. The final feature set was therefore 4 features \times 6 segments. This scheme will be referred to as *passband* bin averaging.

These two schemes – Melscale and passband – will be used as representatives of bin averaging schemes for the transient MES.

4.3.3.4 Relative Performance of Dimensionality Reduction Strategies

In this section, the relative performance of feature selection (CS), feature projection (PCA), Melscale averaging and passband averaging schemes will be evaluated when using the STFT as a basis for feature extraction.

Figure 4.27 provides a direct comparison of the four dimensionality reduction strategies. An original feature set of STFT coefficients has been reduced by each of the methods, and the classification error has been determined on the test dataset at all possible feature dimensions. The results represent the average response across all subjects. Figure 4.27(a) depicts the performance when using a LDA classifier, and (b) shows that when using a MLP classifier.

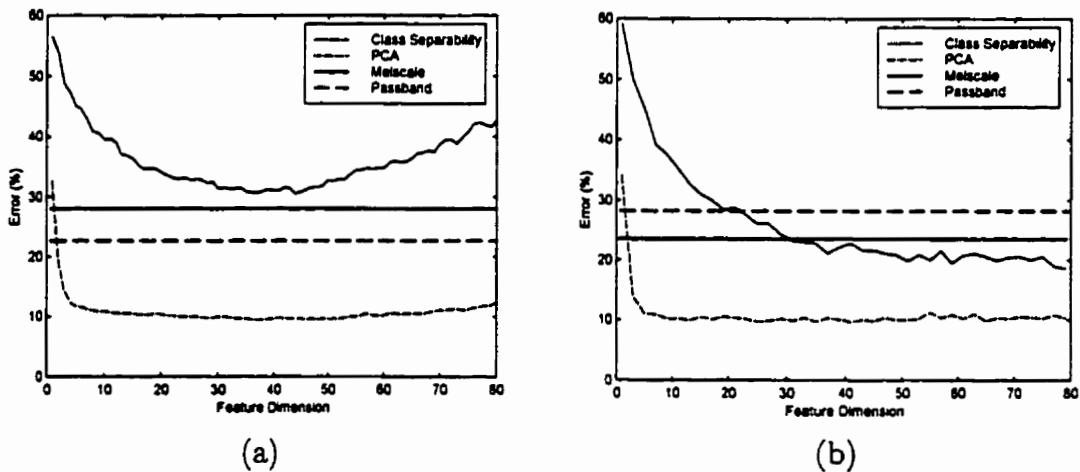


Figure 4.27 – A comparison of dimensionality reduction strategies when using the STFT. The classification error, averaged across all subjects, is shown as a function of feature set dimension. Figure (a) shows the performance using an LDA classifier, with dimensionality reduction performed by feature selection (class separability), feature projection (PCA), Melscale bin averaging, and passband bin averaging. Figure (b) shows the same results when using a MLP classifier.

It is immediately obvious that feature projection dramatically outperforms the other forms of dimensionality reduction. Not only is the average error rate much lower, but this performance is achieved at a very low feature dimension (between 5-10 PCA features). This is due to the dispersive nature of the transient myoelectric signal's STFT. The PCA effectively concentrates the information from the STFT time-frequency cells, amongst which the information may be liberally dispersed. Correspondingly, feature selection performs relatively poorly, especially when using a LDA classifier. This is due to the fact that the LDA is particularly prone to the deleterious effects of the curse of dimensionality. The

dispersion of information amongst the STFT coefficients necessitates a very large dimension to adequately discriminate amongst classes when selecting subsets of individual cells. This large dimension adversely affects the generalization performance of the LDA, given a training set of limited size (100 patterns, in this case). The MLP is less vulnerable to the curse of dimensionality; as a result, the test set error continues to decline with increasing dimension. Never, however, does the performance of feature selection approach that of feature projection.

The averaging schemes fare modestly, with Melscale showing a slight advantage when using a LDA, and passband averaging when using a MLP. Interestingly, when using a LDA, both averaging schemes outperform the more systematic feature selection strategy. This emphasizes the difficulty in isolating discriminatory information in individually selected cells when using feature selection. The averaging schemes actually convey more information (in a limited number of dimensions) by smoothing adjacent cells, albeit in an *ad hoc* manner. With the MLPs ability to handle larger dimensions, feature selection outperforms the averaging schemes.

The results above convey generalized performance with respect to feature set dimension; the performance at each dimension is computed for each subject, and the ensemble average of these responses is then determined. Once again, a more meaningful comparison would be to determine the "optimal" feature dimension for each subject, and compare the ensemble of "optimal" responses. Note that the determination of "optimal" feature set dimension is irrelevant when using averaging schemes.

Figure 4.28 shows a scatterplot of the test set error for each subject at the “optimal” dimension, using each of the dimensionality reduction schemes. Figure 4.28(a) depicts the response using a LDA classifier, and Figure 4.28(b) when using a MLP classifier. Superimposed on each scatterplot is the mean response of each dimensionality reduction method; the mean using a LDA and a MLP are directly compared in Figure (c).

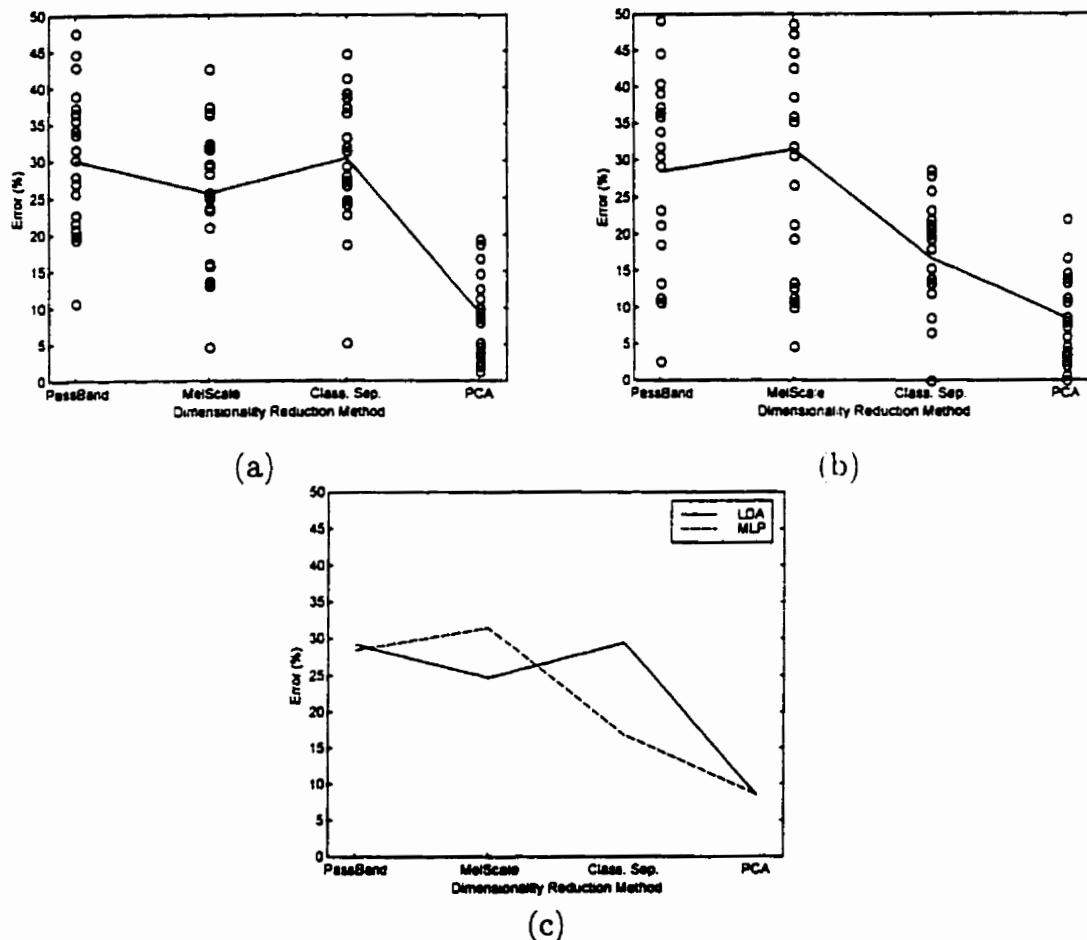


Figure 4.28 – A comparison of dimensionality reduction strategies, using the optimal feature set dimension for each subject. When using feature selection and feature projection, the optimal dimension is estimated using the validation dataset; the error rate is computed from the test set. Figure (a) depicts the performance using a LDA classifier; (b) shows the performance using a MLP classifier; (c) compares the response of the LDA and the MLP, averaged across all subjects.

Again, using the optimal dimensionality for each subject, it is clear that PCA feature projection is the most effective dimensionality reduction scheme. It is

encouraging to observe that the LDA generalizes as well as the MLP when using PCA, implying that this simple classifier can be used effectively with the STFT.

4.3.4 Summary

It has been shown that, when classifying the transient MES, PCA is clearly the most effective means of dimensionality reduction when using a STFT based feature set. The STFT parameters that yield the best classification performance were empirically determined. When using PCA, a hamming window of length 64 and an overlap of 50% gives the STFT its greatest efficacy. The performance of the STFT, relative to other feature sets, will be described in Section 4.6.

4.4 Wavelet Transform Based Features

Within the last ten years, the wavelet transform has developed from a mathematical curiosity into an invaluable tool in signal processing. This section will investigate the efficacy of wavelet transform based features in transient MES classification.

4.4.1 Background

When compared to the enormous interest given to signal analysis, compression and denoising, the wavelet transform has received relatively little attention as a basis for pattern recognition. The wavelet coefficients would seem to constitute a very effective feature set, since the transform tiles the time-frequency plane in a manner that is more appropriate than the STFT for many physical signals. There is a fundamental drawback to the discrete wavelet transform however, in that it lacks *shift-invariance*. If the signal to be analyzed is shifted (even by a small amount), the corresponding wavelet transform coefficients do not experience the same simple translation. Instead, the coefficients are modified in a much more complex manner, due to the fact that the WT is critically sampled¹. Although this is inconsequential when dealing with well-aligned data, it presents a significant problem when classifying signals that are subject to temporal translation. The following are methods that have been devised to compensate for the WT's lack of shift invariance.

¹ A detailed description of the WT's lack of shift invariance is given in Appendix D.

The Continuous Wavelet Transform. Instead of using the WT, one may use the CWT, which preserves the simple relationship between translation in the input signal and translation in the transform [Lin96]. This is because the CWT is not critically sampled, and does not impose any modification upon the transform coefficients, other than the shift. The CWT, however, does not approach the computational efficiency of the WT, making it impractical for real-time applications.

Wavelet Neural Networks. The term “wavelet network” was proposed by Zhang and Beveniste [Zhang92] to describe a feedforward neural network with a single hidden layer, in which the threshold functions of the network’s neurons are replaced by wavelet functions. Whereas these networks were first developed for function approximation, they have been more recently adapted to pattern recognition [Dikhaus96][Szu92][Telfer92][Kadambe94]. Like radial basis function networks, wavelet networks are trained in two stages. First, the appropriate set of wavelet functions (or “wavelons”) must be determined by some criterion. Once the wavelons have been selected, the network is trained using a standard learning algorithm such as backpropagation.

Although wavelet networks have proven useful in some applications, a satisfactory methodology for selecting the number and type of wavelons has yet to be determined. The integration of the feature extraction and classification stages is counterintuitive to the focus here upon signal representation, and therefore wavelet networks have not been considered in this work.

Shiftable Wavelet Transforms. With an appreciation that the lack of translation invariance is a deficiency of the WT, a significant amount of research

has been devoted to developing shift-invariant wavelet transforms. Shift-invariant multiresolution representations exist, but none are completely satisfactory. Some methods employ high oversampling rates [Beylkin92][Saito93], where no down-sampling with the changing scale is allowed. Others require immense computational complexity, in the form of the matching pursuit algorithm [Mallat93]. Another approach has been to give up on shift-invariance, and settle for a less restrictive property termed *shiftability*, which only requires that the energy within each subband remain constant [Simoncelli92].

More recently, several authors have independently proposed orthonormal shift invariant algorithms that involve choosing relative shifts between the basis function of a parent node and its respective children nodes [Pesquet96] [Liang96][Nason95][Coifman95]. At each stage of decomposition, the best circulant shift is selected as that which minimizes an information cost function. This concept has also been extended to wavelet packets [Cohen95][DelMarco94]. Each of these approaches imposes additional computational complexity in the feedforward calculation of the transform coefficients. Cohen [Cohen95] has shown that to be truly shift-invariant, the cost function must descend down to all levels of the decomposition, significantly increasing the complexity.

The best approach to a shift-invariant WT representation is still very much an open issue, and therefore will not be considered in this work.

Shift-Invariant WT Features. Although the WT coefficients are not shift-invariant, it has been shown that the local extrema and the zero crossings of the WT are shift-invariant [Mallat91][Mallat92b]. In fact, it has been shown that the signal can be exactly recovered from the local extrema representation of its wavelet

transform [Mallat92a]. In this approach, the signal is decomposed using a standard wavelet transform, but at each scale only the points which correspond to local extrema (or zero crossings) are retained. A representation by local extrema is somewhat more complete than a zero crossings representation in that the amplitude of the extrema convey additional information. The method of local extrema will be used as a form of feature selection in this section.

Cycle Spinning. The term “cycle spinning” was coined by Coifman *et al.* [Coifman95] to describe the technique of creating several shifted versions of the data to “average out” translation dependence when denoising signals. The lack of translation-invariance produces visual artifacts due to Gibbs phenomenon in the area of discontinuities. Cycle spinning was shown to significantly suppress these artifacts. Further, they showed that denoising when cycle spinning over all circulant shifts is equivalent to denoising using an undecimated wavelet transform. Although there are no published reports of using this technique for pattern recognition, it is reasonable to expect that “bootstrapping” the training set with several shifted versions of the signals would augment generalization. A successful application of this approach to geoacoustic signals has been reported to the author in a personal communication [Saito97]. The effects of cycle spinning upon WT based features will be investigated with respect to the transient MES problem here.

The author believes that a perfect compensation for the lack of translation variance has yet to be offered. Many concur, as was evident from the volley of responses upon querying the Internet journal *Wavelet Digest* [WD5.3 #20, www.wavelet.org]. What has not been considered however, is that the degree to which the lack of translation invariance will affect a given classification problem

depends not only upon the data, but upon the means of dimensionality reduction as well. It will be shown in this section that WT based features derived from feature projection are much less affected by shift than those obtained by feature selection.

4.4.2 Related Work

Having first described the problems associated with using the WT as a basis for pattern recognition, we now turn to the successful applications that might encourage one to apply WT based features to transient MES classification.

If the application does not involve data that is subject to translation, or if the data can be precisely aligned before analysis, then the application of the WT to pattern recognition is fairly straightforward. Examples of these type of applications include image texture segmentation [Bovik92] and classification of MRI spectra [Tate96].

When dealing with signals of biological origin however, the problem of temporal translation is unavoidable. The transient nature of biological signals requires that they be recorded by triggering on the waveform or some external signal. Whether this is done manually or automatically, it is seldom precise. Moreover, many biological signals such as the surface MES have loose temporal structure, making alignment difficult due to structural variations amongst signals.

The problem of characterizing ECG patterns is of enormous interest; it is not surprising that attempts have been made to use wavelet based feature sets. Bentley *et al.* have shown that the WT (using feature selection dimensionality

reduction) outperforms a set of morphological features extracted from a quadratic time-frequency distribution when classifying heart sounds [Bentley98]. No indication was given regarding the methods that may have been used to compensate for temporal translation. Others have successfully used Mallat's local extrema representation to recognize cardiac patterns [Senhadji95][Li95][Gyaw94].

WT based features have been used to perform *detection* of biological events, which is merely a two-class classification problem. The detection of characteristic EEG spikes has been done using subsets of WT coefficients empirically chosen as those responsive to the phenomena of interest [Durka96][Kalayci95]. The detection rate however, still cannot match that of visual analysis by an adept clinician.

Wavelet based features have received considerable attention with respect to classification of underwater acoustic signals. Some have used wavelet coefficients selected using a simple amplitude threshold [Kundu94]. The WT based feature set was shown to outperform a STFT base feature set using this elementary feature selection method, without any compensation for temporal shifts. It has been shown that the performance can be further improved by using more sophisticated dimensionality reduction [Intrator97][Huynh98]. These results are encouraging with respect to the prospect of applying the WT to transient MES classification. The two problems are somewhat similar in the need to detect the onset of a pattern, and the multitude of intrinsic and extrinsic sources of intra-class variance.

To the extent of the author's knowledge, the WT has yet to be used as a basis for MES pattern recognition.

4.4.3 Wavelet Transform Parameter Selection

As compared to the STFT, there are relatively few parameters to select when using the WT. This is because to a certain extent, the time-frequency tiling is predetermined when performing a full dyadic WT decomposition. Only two parameters are of any consequence: the choice of mother wavelet, and the depth of decomposition. The best mother wavelet for transient MES classification will be empirically determined. The WT decomposition may be terminated prior to a full decomposition, resulting in a modification of the low frequency region of the time-frequency tiling scheme. It will be shown that the best classification performance requires a full decomposition.

It should be noted at this point that, in all cases, the energy (the square) of each wavelet coefficient was used instead of the coefficients themselves. The highly irregular waveforms of the transient MES subtend very poor classifier generalization when using signed coefficients. It has been noted in the course of this work that this is not true for waveforms of greater regularity, in which case the signed coefficients are more tightly clustered.

4.4.3.1 The Mother Wavelet

In Chapter 3, the characteristics of the mother (or basic) wavelet were summarized in the following properties:

Vanishing moments: The higher the degrees of vanishing moments a basis has, the better it models the smooth part of the signal.

Regularity: The larger the regularity, the smoother the basis vector becomes.

Compact support: This property is important for efficient and exact numerical implementation. All wavelet families considered here have compact support.

For most wavelet families, the number of vanishing moments and the regularity increases with increasing order (length) of the basic wavelet. An ensemble of the most commonly used mother wavelet functions will be subject to empirical evaluation. These are:

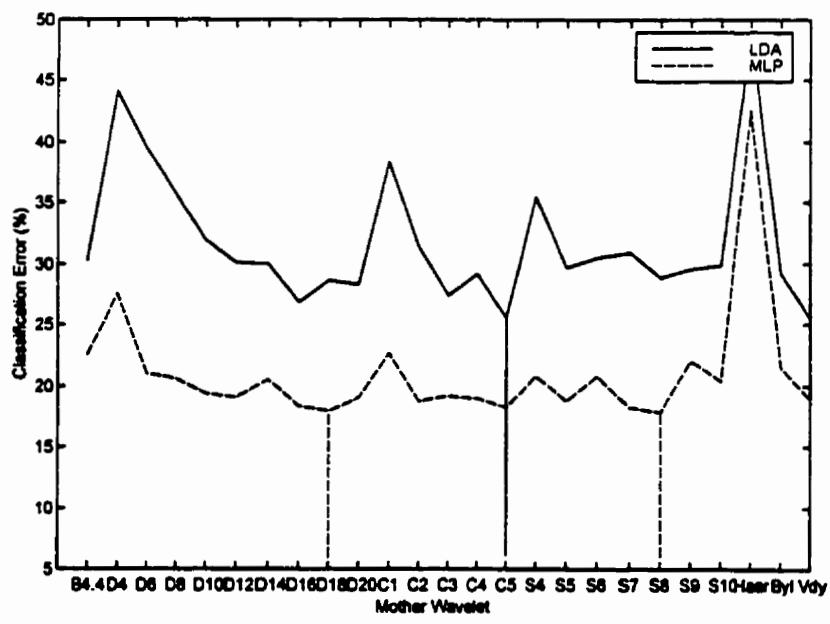
1. **Daubechies wavelets.** These are the original wavelets proposed by Daubechies [Daubechies88]. *Daublets* are compactly supported wavelets that are extremely asymmetric (introducing large phase distortion) and have the highest number of vanishing moments for a given support width. A Daublet of order n (denoted D_n), has a filter length of $2n$, a support width of $2n-1$, and n vanishing moments.
2. **Coiflets.** This family of wavelets was created by Daubechies [Daubechies92] at the request of R. Coifman. They are compactly supported wavelets designed to yield the highest number of vanishing moments for both the mother wavelet and the scaling function for a given width. A Coiflet of order n (denoted C_n), has a filter length of $6n$, a support width of $6n-1$, and $2n$ vanishing moments for the mother wavelet and the scaling function.
3. **Symmlets.** A modification of Daublets, Symmlets are compactly supported wavelets with minimal asymmetry and the highest number of vanishing moments for a given support width. The near symmetry introduces minimal

phase distortion into the transform. A Symmlet of order n (denoted S_n), has a filter length of $2n$, a support width of $2n - 1$, and n vanishing moments.

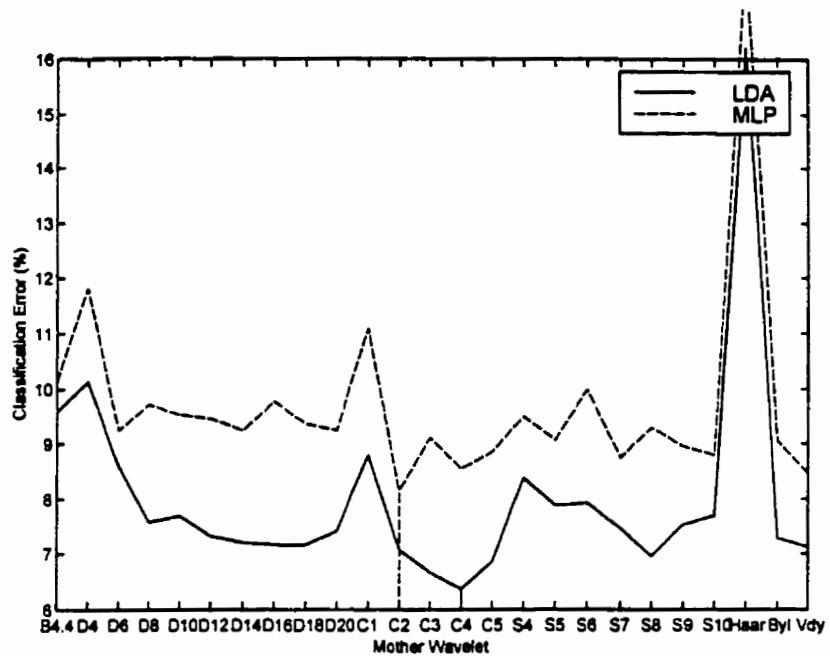
4. **Biorthogonal Wavelets.** By using two wavelets, one for decomposition and the other for reconstruction, perfect symmetry and exact reconstruction are possible. When using a single orthogonal basis, this is not possible [Strang97]. This family of wavelets exhibits the property of linear phase, which is needed for signal and image reconstruction. These wavelets are most often implemented as splines, and are denoted as B_{n_r, n_d} , where n_r and n_d are the order of the reconstruction and decomposition wavelets, respectively. The filter length is $\max(2n_r, 2n_d) + 2$ and the support width is $2n_r + 1$ for reconstruction and $2n_d + 1$ for decomposition. The number of vanishing moments of the mother wavelet is $n_r - 1$ for both reconstruction and decomposition [Cohen90].
5. **The Beylkin filter** places roots for the frequency response function close to the Nyquist frequency on the real axis [Wickerhauser94]. This is denoted here as Byl .
6. **The Vaidyanathan filter** gives an exact reconstruction, but does not satisfy any condition regarding vanishing moments. The filter has been optimized for speech coding. It is denoted here as Vdy .
7. **The Haar filter.** A family of rectangular windows (which could be considered a Daubechies-2) was the first wavelet, though not called as such, and is discontinuous.

The manner in which vanishing moments, regularity and compact support affect the wavelet's efficacy as a basis for signal classification is not clear. One would expect that a wavelet that "looks like" the elemental components of the signals under consideration would be the most appropriate. In the time-frequency plane, this translates into the basis that best localizes the signal energy and clearly distinguishes signals in different classes. This will be determined empirically, based upon the ensemble of two channel data.

For each wavelet family, the test set classification error was computed over the entire subject database. The wavelet coefficients were subject to dimensionality reduction using feature selection (CS) and feature projection (PCA). Figure 4.29 depicts the classification error when using each wavelet, averaged across all subjects.



(a)



(b)

Figure 4.29 – The classification error when using various types of mother wavelet. Figure (a) shows the performance when using CS dimensionality reduction, and (b) when using PCA. The vertical lines indicate the wavelet of minimum error for each classifier.

As one would expect, the classification error is much lower when using feature projection, regardless of the wavelet used. The most obvious effect is the relatively poor performance of the wavelets of very low order (short filter length): the D4, C1, S2, and Haar wavelets. Clearly, the low regularity and few vanishing moments do not subtend an appropriate basis for pattern recognition. An interesting trend occurs within the Daublet, Coiflet and Symmlet families. As the filter order grows, the classification performance tends to improve up to a certain point, and then degrade. This implies that a wavelet of considerable (but not maximal) smoothness is that which is best for classification.

Although no single wavelet can be identified as clearly superior, some generalizations can be determined in the context of the method of dimensionality reduction and the classifier. Within the Daublet, Coiflet and Symmlet families,

the best performance results when the filter length is roughly 18, 4 and 8, respectively. The Beylkin and Vaidyanathan filters experience moderate performance. The best performance when using CS comes with the D18 and S8 wavelets. The best overall performance results when using a C4 wavelet and PCA dimensionality reduction with a LDA classifier².

4.4.3.2 The Depth of Decomposition

Aside from the selection of mother wavelet, the only other adjustable parameter when performing a WT is the depth of decomposition. For a signal of length N , the maximum depth of decomposition is $J = \log_2 N$. Recall that the first level of decomposition partitions the frequency axis into high and low subbands. The next level partitions the lower subband into high and low subbands, and so on. A full decomposition proceeds until subdivision is no longer possible: the lowest level detail and approximation subbands are scalars. If decomposition is terminated before a full decomposition, the frequency axis is not fully partitioned toward zero. In this case, the approximation signal occupies the unpartitioned portion of the frequency axis. A full and partial WT decomposition are shown in Figure 4.30.

It is obvious that a partial WT decomposition trades off frequency resolution for temporal resolution. This may improve or worsen the efficacy of the WT feature set as a basis for classification, depending upon the nature of the low frequency region of the transient MES.

² For the purposes of comparative analysis when using the WT, a D18 wavelet will be used when using CS, and a C4 wavelet will be used when using PCA.

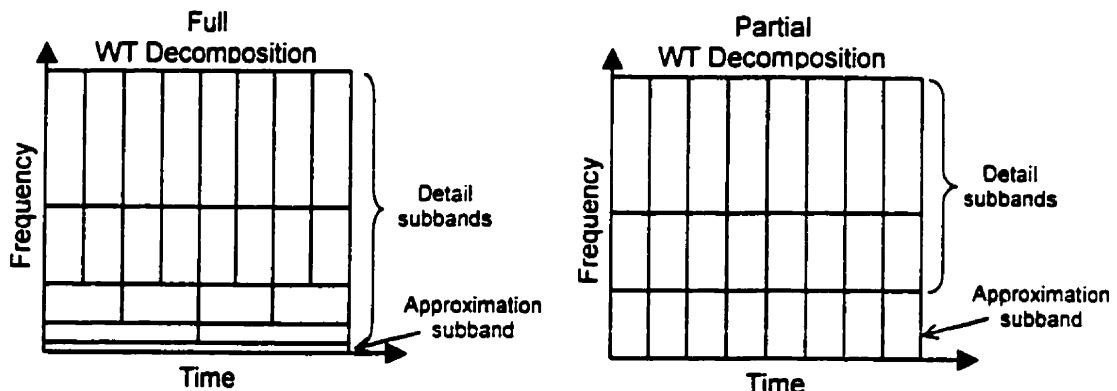


Figure 4.30 – The time-frequency tiling of a full and partial WT decomposition.

For each subject, a WT feature set was determined from WT decompositions ranging from a single level to a full decomposition ($J = \log_2 256 = 8$). The test set classification error averaged across all subjects, using CS and PCA dimensionality reduction, is shown in Figure 4.31.

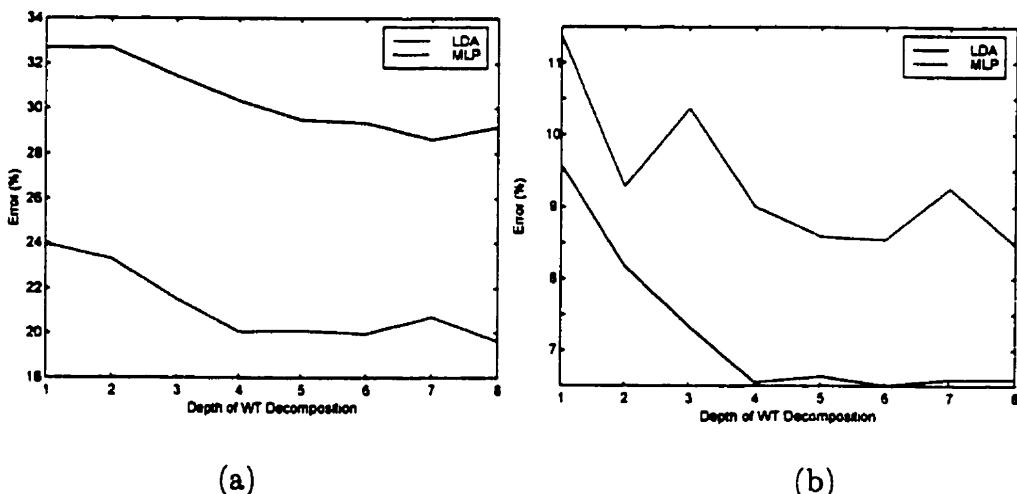


Figure 4.31 – The effect of the level of WT decomposition upon classification error, averaged across all subjects. Figure (a) shows the results when using CS, (b) when using PCA dimensionality reduction.

Regardless of the classifier or dimensionality reduction scheme, the classification error improves with increasing depth of decomposition. It is clear that the frequency resolution in the lowest subband provided by a full WT decomposition is important for transient MES classification.

4.4.4 Dimensionality Reduction

In Section 4.3 it was shown that an appropriate form of dimensionality reduction is crucial to the success of the STFT as a basis for classification. The dimension of the WT of a signal of length N is also N . For the transient MES, there are two channels of length $N = 256$, yielding 512 wavelet coefficients which describe a single pattern. The role of dimensionality reduction when using the wavelet transform is examined here. As before, feature selection (using class separability) and feature projection (using PCA) are among the candidate techniques. Two other methods that have been used to derive features from a wavelet decomposition will be examined: a representation by wavelet transform *local extrema*, and a representation by wavelet transform *subband energy*. These methods will be described in the following sections.

4.4.4.1 Feature Selection

A reduction of the WT using feature selection has been performed here using a class separability measure. As with the TD and STFT representations, add-on and knock-out methods did not demonstrate sufficient merit to warrant their additional computational expense. As determined in the previous section, the D18 wavelet yields the best generalization performance when using CS, and therefore this wavelet will be used whenever using CS.

To evaluate the efficacy of CS when using a WT based feature set, consider the classification error of the validation set at all possible feature set dimensions.

Figure 4.32 depicts this response for the WT, as compared to that of the TD and STFT based representations.

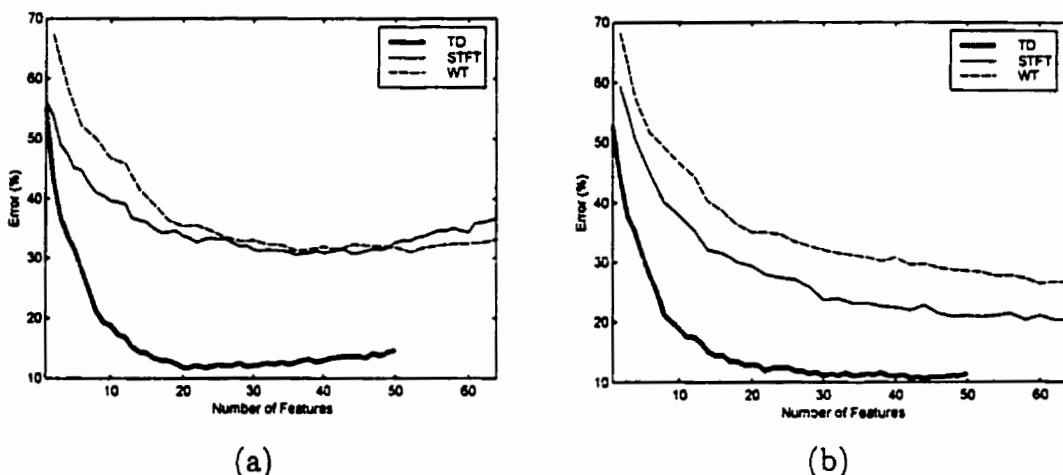


Figure 4.32 – The effect of dimension when using CS-reduced TD, STFT and WT feature sets.
Figure (a) depicts the validation set error, averaged across all subjects when using a LDA classifier, and (b) when using a MLP classifier.

As was observed with the STFT, a CS-reduced WT feature set performs very poorly, as compared to the TD feature set. This is due to the inability of feature selection methods to concentrate sufficient information in a feature subset of reasonable dimension. It is clear that a WT basis performs no better than a STFT basis, even slightly worse when using a MLP. The lackluster performance of the WT may be due to the fact that the STFT's time-frequency tiling is more appropriate than that of the WT. This is possible, but unlikely, since the time-frequency signature of the transient MES would seem to benefit from the nature of the WT's coverage.

A more likely possibility is that the WT coefficients are subject to nonlinear modification due to their lack of shift invariance. The manner in which the patterns are triggered introduces a variable offset, introducing an additional source of intra-class variance into the WT coefficients. The mechanism of feature

selection has no means of compensating for shift-related modification, since the reduced feature set is merely a subset of the WT coefficients.

4.4.4.2 Feature Projection

In the previous section, the C4 wavelet was shown to offer the best generalization performance when using PCA; this will be the wavelet of choice here. Feature projection using PCA has been shown to be much superior to CS methods when using the STFT. To evaluate the efficacy of PCA when using a WT based feature set, consider the classification error of the validation set at all possible feature set dimensions. Figure 4.33 depicts this response for the WT, as compared to that of the TD and STFT based representations.

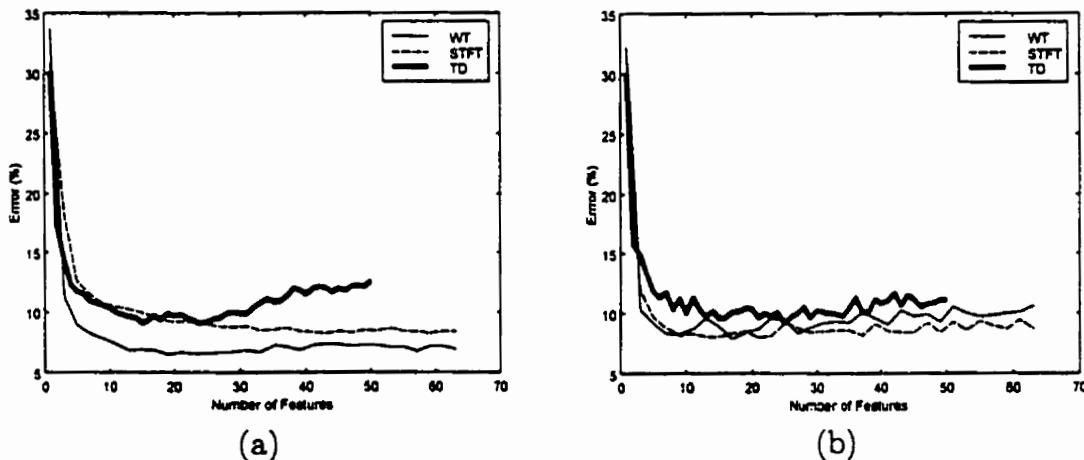


Figure 4.33 – The effect of dimension when using PCA-reduced TD, STFT and WT feature sets. Figure (a) depicts the validation set error, averaged across all subjects of the feature sets when using a LDA classifier, and (b) when using a MLP classifier.

When using PCA dimensionality reduction, the WT and STFT outperform the TD feature sets. When using a LDA classifier, the WT feature set outperforms that based upon the STFT. When using a MLP, there does not seem to be any significant difference between the two. Not only does PCA improve the overall performance of the WT feature set as compared to CS, but it also improves its

performance relative to the STFT. This indicates that the PCA dimensionality reduction is compensating for the shift-related modification of the WT coefficients. Although the lack of shift invariance may introduce additional intra-class variance into a WT TFR, if the class boundaries remain substantially linear, the PCA can consolidate the dispersed features. This appears to be the case, and PCA seems to accommodate, to some degree, the modifications due to temporal shift. The effects of temporal translation upon the WT, and PCA's ability to alleviate these effects are explained in Appendix D.

4.4.4.3 Wavelet Transform Local Extrema

Mallat *et al.* have demonstrated that, although the WT coefficients suffer from nonlinear modification due to signal translation, the local extrema and the zero crossings of the WT are shift-invariant [Mallat91]. When using this approach in image characterization, the signal was decomposed using a standard wavelet transform, but at each scale only the points which correspond to local extrema (or zero crossings) were retained [Mallat92b]. This approach may be viewed as a form of feature selection, such that only a subset of the coefficients are retained according to some *a priori* knowledge of their importance.

A representation by local extrema is somewhat more complete than a zero crossings representation in that the amplitude of the extrema convey additional information. When using the method of WT local extrema, one must include an index to denote the location of the extrema within each subband. Therefore, the dimensionality of the representation is $2 \times LE$, where LE is the number of local extrema in the WT. When this method of signal representation is applied to

transient MES patterns, it is immediately obvious that there are a large number of local extrema, due to the highly irregular structure of the patterns. It is not uncommon, in a record of 256 points, to observe as many as 150 WT local extrema. In a two channel recording, this results in a feature dimension of $2 \text{ channels} \times 2 \times 150 \text{ extrema} = 600$. This does not result in an effective form of dimensionality reduction, and is obviously too high to be useful³.

A slightly different approach as been suggested by Senhadji *et al.* [Senhadji95]. In proposing a method of cardiac signal recognition, they successfully applied a local extrema representation that retains only one extremum in each WT subband. In a WT decomposition of J levels, this specifies $J+1$ extrema (including the approximation subband) and $J-1$ indices (no indices are needed for the lowest level detail and approximation subbands, which are scalar quantities), yielding a feature set dimension of $2J$. For the two channel transient MES with a record length of $N = 256$, this yields a feature set dimension of $2 \times 2 \times \log_2 256 = 32$.

The performance of this method is given in Section 4.4.4.5, where it is compared to the other methods of WT dimensionality reduction considered here.

4.4.4.4 Wavelet Transform Subband Energy

If it is presumed that a substantial degree of temporal dispersion is present in a set of signals, one may choose to smooth the WT by computing the energy in each

³ Indeed, this method must be followed by CS or PCA to achieve satisfactory performance. When using local extrema followed by CS, a slight improvement to CS alone was observed. When using local extrema followed by PCA, the results were slightly worse than PCA alone. This indicates that processing by local extrema

WT subband. This approach has been used in the classification of underwater mammal sounds [Learned95]. Clearly, discarding all temporal information is a drastic approach to accommodating the effects of temporal shift. When applied to the $N = 256$ records of two channel MES, a feature set of $2 \times 9 = 18$ energy coefficients results. The performance of this method will be compared to the other forms of dimensionality reduction in the next section.

4.4.4.5 The Relative Performance of Dimensionality Reduction Methods

Having described four methods of providing dimensionality reduction for WT coefficients, their relative performance will now be described. The data from each subject were processed using a WT. These coefficients were then subject to dimensionality reduction by CS, PCA, local extrema (LE), and subband energy (SE). When using CS and PCA, the reduced feature set dimension was selected as that which minimized the validation set classification error. The feature set size when using LE and SE methods was 32 and 18, respectively.

The D18 wavelet has been shown to yield the best response when using CS; LE is a form of feature selection and therefore, will use the D18 wavelet as well. SE involves a linear combination of coefficients, and therefore, will be applied using the C4 wavelet which has been determined to perform the best when using PCA analysis. Figure 4.34 shows the test set classification error for each subject when using each of the methods.

performs a form of intelligent feature selection, but that it compromises the performance of PCA by discarding

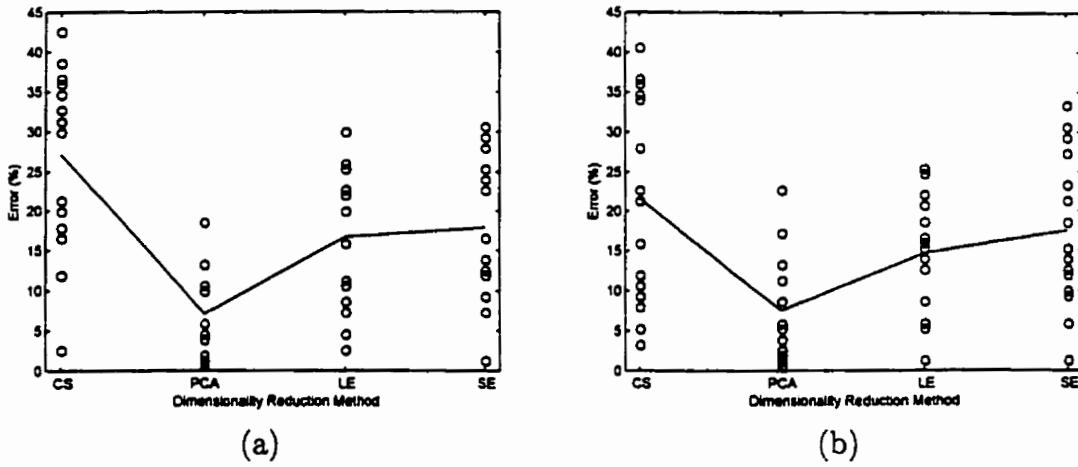


Figure 4.34 – The relative performance of the WT when using CS, PCA, LE and SE based dimensionality reduction methods. Figure (a) depicts the results when using a LDA classifier, and (b) when using a MLP.

The response when using the two classifiers is very similar. The performance is clearly superior when using PCA dimensionality reduction, and the poorest results occur when using CS. The fact that representation by SE outperforms CS is interesting; averaging all temporal information yields better results than selecting the best subset of time-frequency coefficients. This implies that the effect of temporal dispersion (due to simple translation and due to nonlinear modification due to the lack of shift-invariance) is so severe in the WT so as to warrant discarding temporal information in lieu of using it for classification.

Like the SE representation, the method of LE is immune to modification due to the lack of shift-invariance, as it retains only the points that are invariant. Its advantage over SE is that it encodes temporal information *via* the indices of the local extrema. This extra information allows the LE feature set to outperform the SE representation.

some useful information.

It may be concluded that that temporal information is important for classification, and that the normal temporal dispersion due to misalignment of the patterns can be accommodated. It is the effects of nonlinear modification due to the lack of translation that severely hamper a feature set. This is why the LE method of dimensionality reduction fares better than CS. Overall however, PCA provides superior performance to all other methods. By projecting the WT coefficients onto the orthogonal axes of maximum variance, the PCA accommodates dispersion in time and in frequency. As well, PCA produces an uncorrelated feature set, yielding a parsimonious signal description which aids the performance of the LDA classifier.

4.4.5 Summary

It has been shown that, when classifying the transient MES, PCA is clearly the most effective means of dimensionality reduction when using a WT based feature set, as was the case with the STFT. The choice of mother wavelet does not dramatically affect the classification performance, except that wavelets of very short filter length produce relatively poor results. The wavelet functions that provided consistently good performance were the D12-18, C4-5, and Sym5-8 families. Although it is possible to perform a partial WT decomposition, it has been shown that a full WT decomposition produces the best classification results. The performance of the WT, relative to other feature sets, will be described in Section 4.6.

4.5 Wavelet Packet Based Features

The wavelet packet transform (WPT) is a relatively recent generalization of the wavelet transform. A full WPT decomposition provides an overcomplete set of orthogonal bases, amongst which the best may be chosen to suit a particular application. As described in Section 3.4, the “best basis” algorithm was originally proposed for signal compression [Coifman92]; it prunes the overcomplete wavelet packet basis, seeking the best orthonormal basis according to an entropy-based cost function.

It has been suggested however, that the best basis for classification may be very different than that considered optimal for signal compression. In his 1994 Ph.D. dissertation [Saito94], Naoki Saito developed what he termed the *Local Discriminant Basis* (LDB) algorithm, and showed its utility in classifying geoacoustic waveforms. The LDB method is a modified pruning algorithm that uses a cost function that provides a class separability index [Saito95]. The LDB algorithm will be the primary means of constructing wavelet packet bases in this work.

4.5.1 Related Work

The main advantage of the WPT over the WT is its ability to specify a time-frequency tiling that may be adapted to a particular need. Whereas the octave-band tiling of the wavelet transform has been shown to be suitable for many physical signals, it is seldom optimal, and sometimes inappropriate. The applications that have most demonstrably benefited from the flexibility of the WPT have been related to signal compression and denoising.

Researchers at Yale University were asked by the U.S. Federal Bureau of Investigation to help in improving the results on the commercial JPEG standard for fingerprint image compression, which was about 5:1 before the images were unacceptably distorted. Wavelet packet compression has allowed a 20:1 compression with the same distortion standard [Bruce96]. The utility of wavelet packet methods has also been shown as a mechanism of noise removal, providing an effective basis for separating signal and noise subspaces [Saito94b][Kuehner97].

The use of the WPT as a basis for pattern recognition however, has been a most recent development. The first use of the WPT in this manner appears to be as a basis for transient signal detection [DelMarco94], a much simpler task than generalized pattern recognition. This detector used a matched filter; the wavelet basis was selected as that which provided the best basis (least entropy) for modeling the template signal.

A different approach was taken by Learned *et al.* [Learned95]. They did not attempt to choose a single orthonormal basis, but rather, applied a singular value decomposition to a structure comprising the energy map of each wavelet packet subband. By determining the eigenvectors of greatest significance, and selecting the dominant coefficients from the first eigenvector, they compiled a feature set for classifying underwater mammal sounds. This energy map averages the coefficients within each subband however, thereby discarding all temporal information. Only a modest classification performance was possible with this method.

Saito's LDB algorithm [Saito94a][Saito95] has been the method that has seen the greatest acceptance. In the original LDB algorithm, a class separability criterion

applied to the wavelet packet energy map is used to construct the “best” basis for classification. Saito has since provided an augmentation to the algorithm which uses empirical probability density functions instead of energy maps [Saito96a] which, in some cases, provides a better generalization of class separability. Saito’s LDB has been applied to geoacoustic signal classification [Saito96b], radar signal classification [Guglielmi96], underwater mammal sounds and backscattered signals [Huynh98][Intrator97a], and classification of neuron firing patterns in monkeys [Warner96].

Buckheit and Donoho [Buckheit95] have proposed an algorithm that is similar to Saito’s. The method – called *discriminant pursuit* – measures the discrimination power of each basis function using a one-dimensional Fisher’s discriminant. It has been demonstrated, however, that this method is particularly prone to the curse of dimensionality due to the fact that some of the WPT basis functions may be sparsely represented for many sets of signals [Intrator97b].

As was done with TD, STFT, and WT feature sets, the transform parameters and the dimensionality methods which best suit the WPT will be determined in the context of the transient MES classification problem.

4.5.2 Wavelet Packet Transform Parameter Selection

In Section 4.4.3, it was explained that the parameters of the WT include the choice of mother wavelet and the depth of decomposition. When using the WPT, there is an additional transform parameter: the method used to determine the best

orthogonal basis from the full WPT decomposition. As indicated previously, the LDB algorithm will be the method used to specify the WPT bases here. An improvement to the LDB is proposed, based upon segmenting the record to obtain temporally localized LDBs. This is motivated by the fact that the data are obviously nonstationary.

4.5.2.1 Selection of the Mother Wavelet and LDB Cost Function

The choice of mother wavelet and the basis selection cost function will be based upon the performance of the two channel transient MES dataset. They will be treated simultaneously, since the best wavelet must be determined in the context of the means of basis selection.

As described in Chapter 3, the LDB algorithm prunes the WPT tree by evaluating the importance of each basis function. This importance, in the context of classification, is conveyed by means of a class separability measure. The measures (or cost functions) that have been proposed include I -divergence (relative entropy), J -divergence (symmetric relative entropy), and Euclidean distance [Saito94]. Refer to Section 3.4.2.1 for details.

For each LDB cost function, the performance of each mother wavelet was determined, for each subject. For the sake of brevity, only the PCA-reduced feature sets will be considered in this section, since they substantially outperform the CS-reduced sets. Figure 4.35 shows the test set classification error, averaged across all subjects, when using an I -divergence cost function. Here, the feature set dimension has been determined as that which minimizes the validation set error.

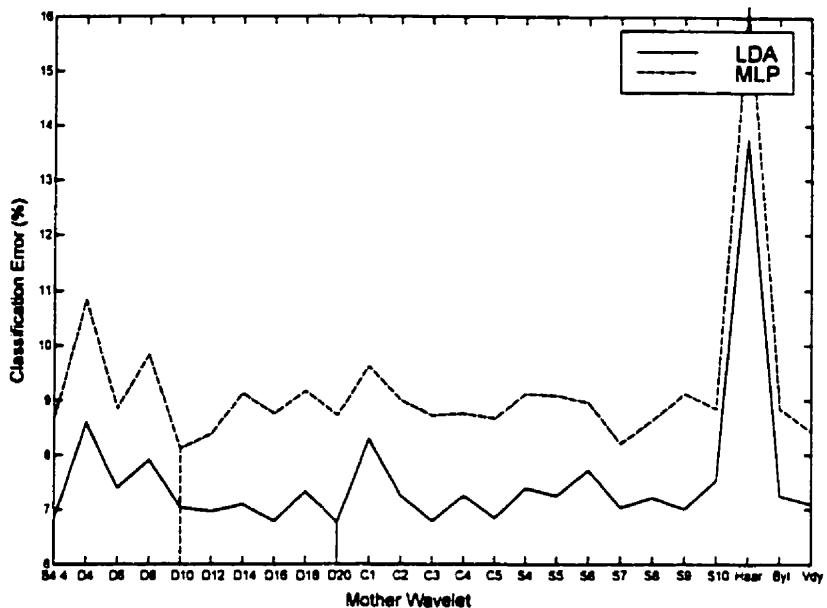


Figure 4.35 – The test set classification error of various mother wavelet families, when using an I-divergence LDB cost function. The vertical lines indicate the mother wavelet subtending the lowest error for each of the LDA and MLP classifiers.

Most notable is that the LDA classifier outperforms the MLP, as was the case when performing PCA upon WT coefficients. This implies that the class boundaries in the high dimensional WPT representation are substantially linear. As with the WT, the higher-order Daubechies, Coiflet, and Symmlet wavelets subtend the best performance. Figure 4.36 depicts the same analysis, repeated using a *J*-divergence cost function.

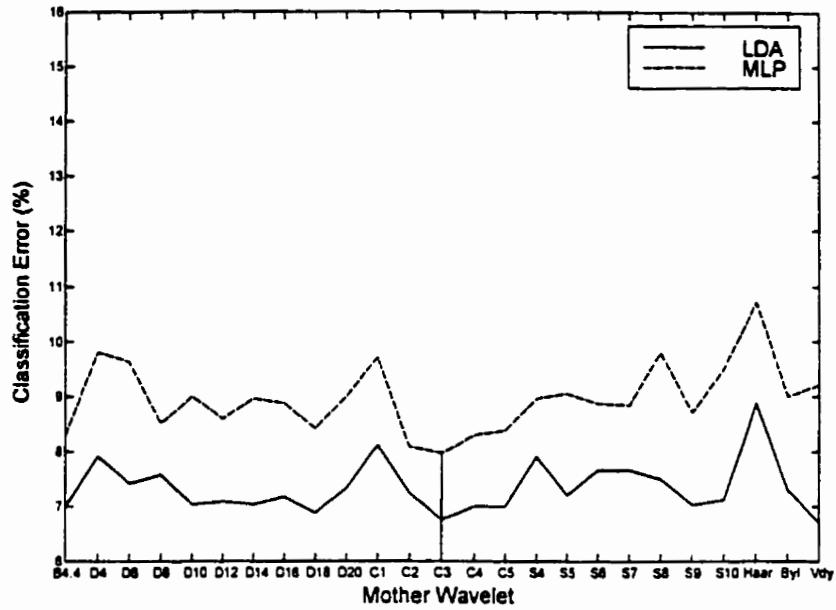


Figure 4.36 – The test set classification error of various mother wavelet families, when using a J-divergence LDB cost function.

The same trends are evident, although the method is more forgiving of some wavelets that perform exceptionally poorly when using I-divergence. The last cost function considered for the LDB algorithm is that of Euclidean distance. The results when using this method are shown in Figure 4.37.

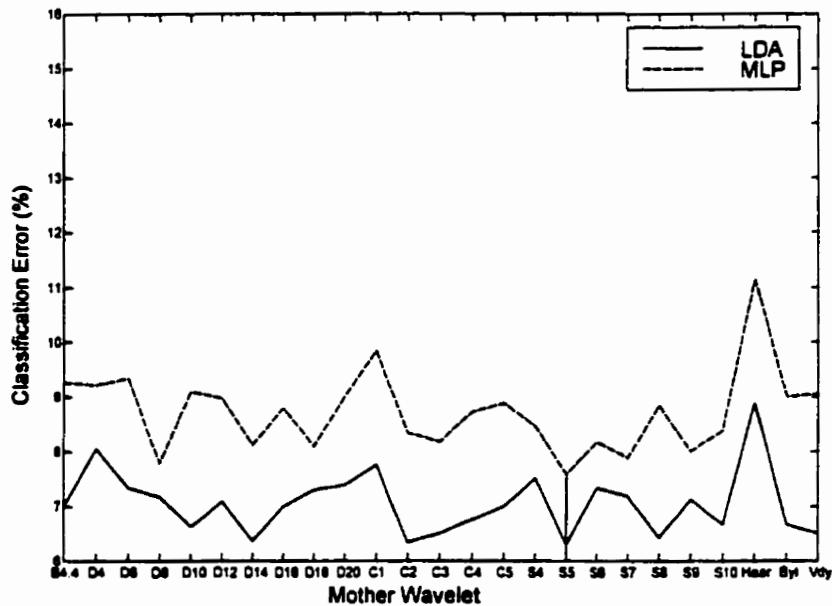


Figure 4.37 – The test set classification error of various mother wavelet families, when using a Euclidean distance LDB cost function.

Although it has been indicated that the LDB algorithm is the primary means of basis selection here, a comparison was done using a modified form of Coifman and Wickerhausers' Best Basis algorithm, which is intended for signal compression. The original Best Basis algorithm determines the best basis by applying an entropy cost function to each individual pattern to be compressed [Coifman92]. To provide a basis for signal classification however, the same basis must be used for all patterns. Therefore, the full WPT was averaged across all patterns in the training set, and the best basis algorithm was applied to this average wavelet packet expansion. What results is a basis that can be considered the best generalized orthonormal basis for reconstructing all patterns in the training set. This cost function is clearly not optimal for class discrimination, but the measure of the importance of each basis function is very stable, since it is derived from the averaged WPT. The test set error for the ensemble of wavelet families when using this entropy cost function is shown in Figure 4.38.

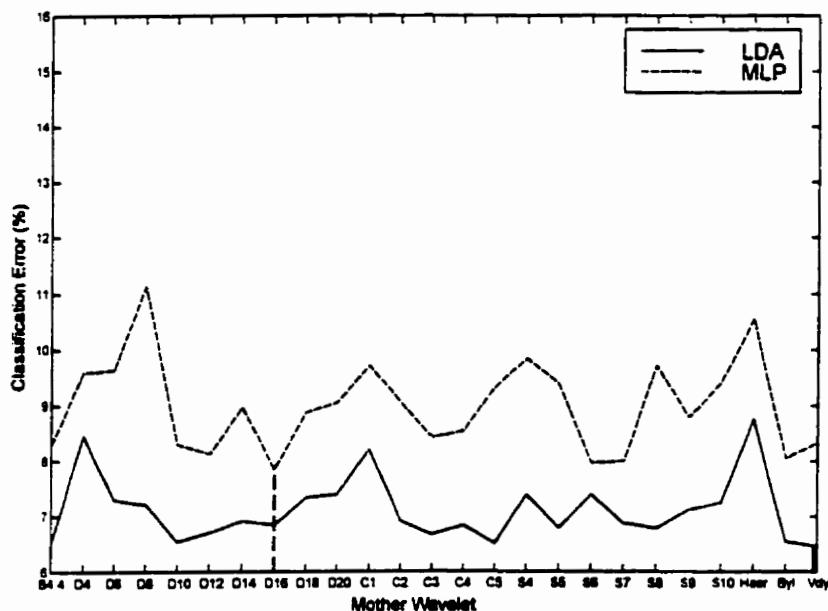


Figure 4.38 – The test set classification error of various mother wavelet families, when using the compression-based Best Basis algorithm with an entropy cost function .

This compression-based pruning algorithm appears to subtend a classification performance that is very similar to the LDB-based strategies.

To provide a direct comparison between performance of each cost function, the classification error corresponding to the *minimum* and the *mean* of each method was determined. Figure 4.39 provides this comparison.

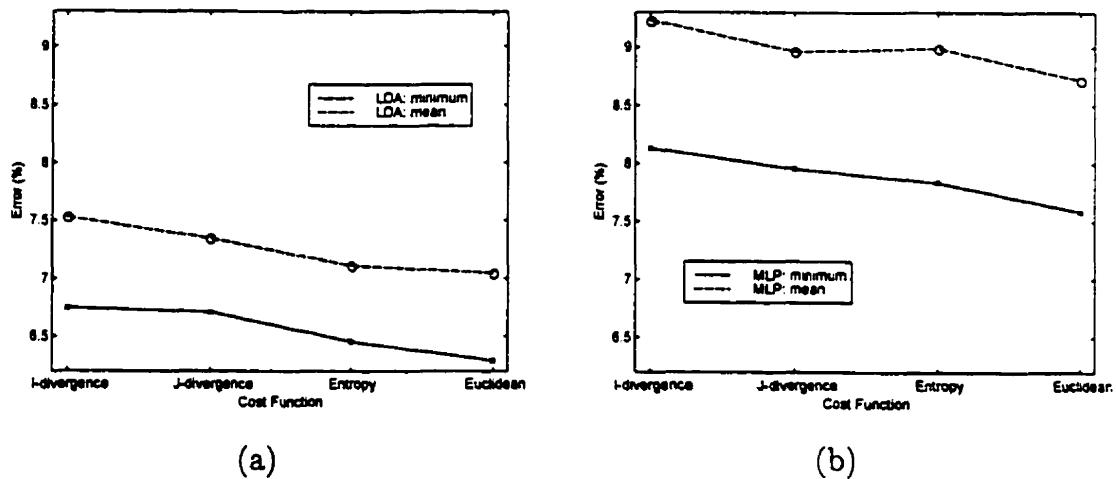


Figure 4.39 – The relative performance of each basis selection cost function, including the LDB class separability indices of I-divergence, J-divergence and Euclidean distance. Also shown is the response of the compression-based Best Basis algorithm which uses an Entropy measure. For each method, the response is shown for the mean error across all wavelet types and for the wavelet subtending the minimum error. Figure (a) depicts the results when using a LDA classifier, and (b) when using a MLP.

For both classifiers, and for both the minimum and mean response, the performance improves as one progresses from I-divergence to J-divergence to Entropy to Euclidean distance¹. The most interesting feature is that the compression-based Best Basis algorithm outperforms two of the LDB algorithms. Although the LDB methods attempt to specify a basis that is more appropriate for classification, the use of a class separability index implies the knowledge of class

¹ The order of presentation of the methods has been chosen to show this progression.

membership. The use of class membership introduces a tendency to bias the evaluation of importance of each basis function to the training set. Therefore, although class separability will yield a basis that is adept at discriminating the training set, it may not generalize well to the test set (or validation set). The class separability information provided by Euclidean distance measure seems to overcome the bias of the LDB algorithm however, as the Euclidean-based LDB provides the best performance overall.

To summarize, the best performance is achieved when using the LDB algorithm with an Euclidean distance measure of class separability. The mother wavelet that yields the lowest classification error in this case is the Symmlet-5 wavelet, for both the LDA and MLP classifiers. As was the case with the WT, the LDA is clearly superior to the MLP when using PCA-reduced WPT features, presumably due to inherently linear class boundaries.

4.5.2.2 A Temporally Segmented LDB

The WPT involves a decomposition into an overcomplete set of frequency subbands. From this overcomplete subband representation, an orthonormal basis is chosen according to some criterion. We may say then, that the orthonormal basis specifies the best segmentation of the *frequency axis* with respect to the chosen criterion. Each subband in this partition has a time resolution inversely proportional to its bandwidth, as a consequence of the critically sampled time-frequency grid.

This optimization with respect to frequency however, is with respect to the entire record. If the frequency characteristics of the signals are nonstationary, it is possible that a temporally localized optimization with respect to frequency may produce a more appropriate time-frequency tiling. Figure 4.40 depicts a temporally segmented LDB representation.

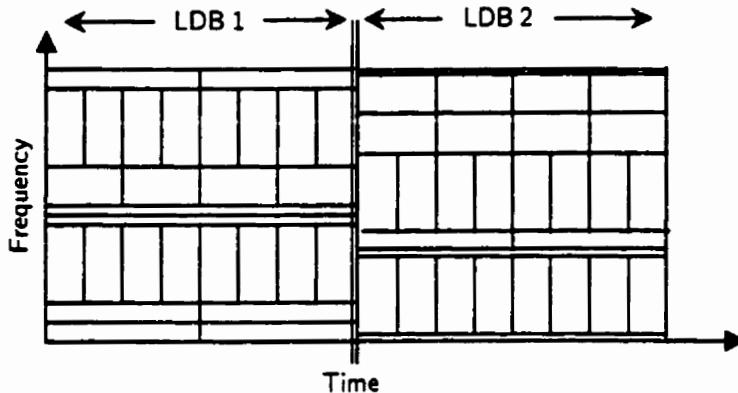


Figure 4.40 – The time-frequency tiling of a temporally segmented LDB representation. Two segments are shown in this example.

The frequency characteristics of the transient MES are clearly nonstationary, motivating the use of a temporally segmented LDB. The records of two channel MES were divided into segments of equal size, and a LDB was determined with respect to each segment. The WPT coefficients were computed from each segment and concatenated to form an aggregate set of features, which were then subject to both CS and PCA dimensionality reduction. This procedure was done with one, two, four and eight segments. Figure 4.41 depicts the test set classification error using these segmentation schemes.

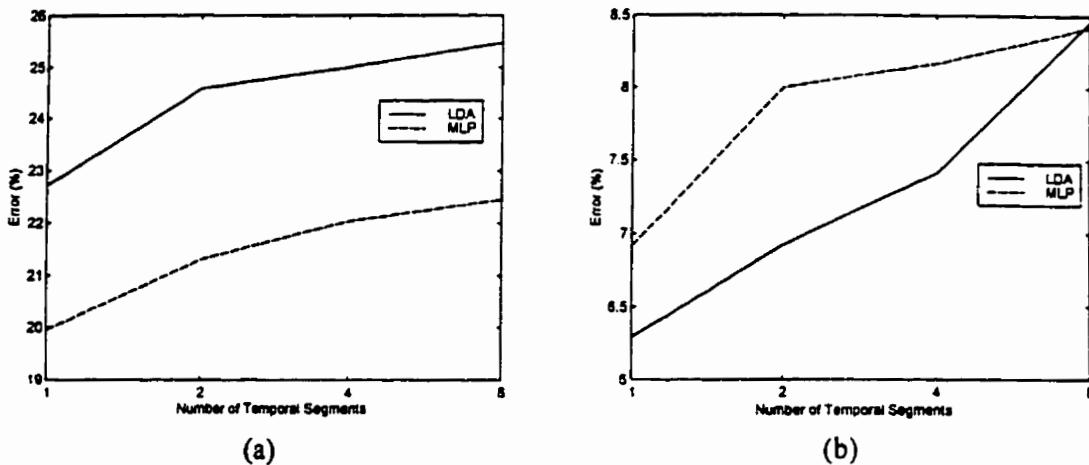


Figure 4.41 – The test set classification error of a temporally segmented LDB, averaged across all subjects. Figure (a) shows the effect of temporal segmentation when using CS, and (b) when using PCA.

Clearly, temporal segmentation does not have the desired effect. The likely reason for this is that, by limiting an LDB analysis to a subset of the entire record, one sacrifices the resolution of the narrowest frequency band that may be specified. For a signal of length n samples, the WPT may be decomposed to $J = \log_2 n$ levels, such that the narrowest subband is $2^{-J} = \frac{1}{n}$ times the range of the frequency axis. If the record is segmented by four however, the narrowest subband becomes $\frac{1}{(n/4)} = \frac{4}{n}$, decreasing the frequency resolution by a factor of four. This apparently has a deleterious effect upon the efficacy of the feature set for classification of the transient MES, so much so that it outweighs the benefits of localizing the frequency optimization. Therefore, no temporal segmentation is desirable when using the LDB. The effects of compromising frequency resolution were also observed when terminating a WT before full decomposition, in Section 4.4.3.2.

4.5.3 Dimensionality Reduction

It has become very clear that the success of the STFT and the WT as feature sets for classifying transient MES patterns depends resolutely upon the means of dimensionality reduction. There is no reason to expect that the WPT should behave differently. This section will conclude the investigation of the effects of dimensionality reduction upon TFR based feature sets.

The WPT was applied to the data of each subject in the transient MES database. The LDB algorithm with an Euclidean distance cost function was used to determine the basis, and the Symmlet-5 wavelet family was used. This combination was shown to yield the best performance upon these data in the previous section.

The WPT coefficients were subject to CS feature selection and PCA feature projection. Figure 4.42 shows the classification error of the validation set when using a CS-reduced WPT feature set. The averaged response across all subjects is shown as a function of feature set dimension. The analogous performance of the TD, STFT, and WT features are shown for comparative purposes.

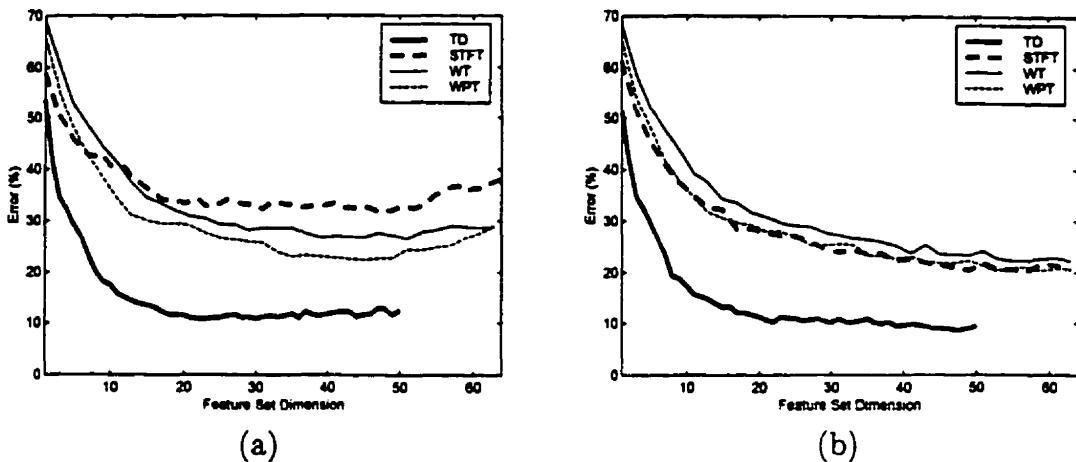


Figure 4.42 – The effect of feature set dimension when using CS-reduced TD, STFT, WT and WPT feature sets. Figure (a) depicts the validation set error, averaged across all subjects when using a LDA classifier, and (b) when using a MLP classifier.

When using a LDA classifier, the WPT demonstrates superior performance to the other TFR feature sets, albeit much worse than the TD set. When using a MLP classifier, the WPT subtends essentially the same response as the STFT. In both cases however, the WPT outperforms the WT, as would be expected. The WPT is vulnerable to the effects of the lack of translation invariance in the same manner as the WT, and this has a deleterious effect on the generalization ability of the feature set. The advantage of the WPT over the WT is its ability to adaptively specify a time-frequency tiling via the LDB algorithm. This explains the superiority of the WPT to the WT and evidently, is sufficiently advantageous to improve upon the performance of the STFT, which is shift invariant².

The same analysis was done using PCA feature projection. The response is shown in Figure 4.43.

² In the sense that temporal shifts in the signal correspond to simple shifts in the STFT.

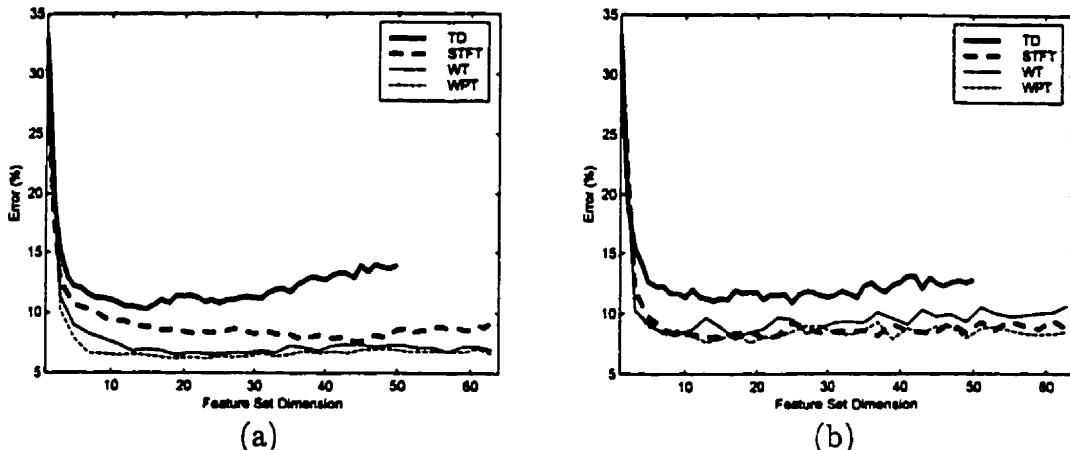


Figure 4.43 – The effect of feature set dimension when using PCA-reduced TD, STFT, WT and WPT feature sets. Figure (a) depicts the validation set error, averaged across all subjects when using a LDA classifier, and (b) when using a MLP classifier.

As witnessed previously, the TFR feature sets enjoy an advantage over the TD set when using PCA feature projection. Here, the WPT demonstrates a slightly better performance than the other TFR feature sets when using a LDA classifier. The TFR feature sets exhibit essentially the same performance when using a MLP classifier. As well, the performance of the WT and WPT when using a LDA is superior to all feature sets when using a MLP.

The plots of validation set performance given here provide valuable information about the interrelationships between classification error, feature sets, feature set dimension and classifiers. The ultimate measure of performance however, is the classification error of the test set, evaluated at the dimension that minimizes the validation set error for each subject. A comprehensive comparison of test set performance between feature sets will be given in the next section.

4.6 Performance Summary

4.6.1 The Relative Performance Amongst Feature Sets

The previous sections have described the application of TD, STFT, WT and WPT based feature sets to the task of transient MES classification. Closely coupled with the use of each feature set is the means of dimensionality reduction. When using TFR based representations, it is crucial that an appropriate form of dimensionality reduction be performed. It has been demonstrated that PCA feature projection dramatically outperforms feature selection methods when using TFR based feature sets, and marginally so when using TD features.

Each feature extraction method has parameters that may be specified, such as temporal segmentation (when using TD), windowing methods (when using STFT) or wavelet bases (when using WT, WPT). Parameter selection must be performed in the context of each candidate dimensionality reduction method. Based upon evidence derived from the roster of 16 subjects, the best parameters for each feature set were specified. This section describes the relative performance of each feature set in terms of test set classification error. In each case, the best set of parameters for each feature set will be used, as determined in the context of either CS or PCA. The feature set dimension is selected as that which minimizes the validation set classification error.

Figure 4.44 depicts the test set classification error, averaged across all subjects. Figure (a) shows the response when using CS, and Figure (b) when using PCA.

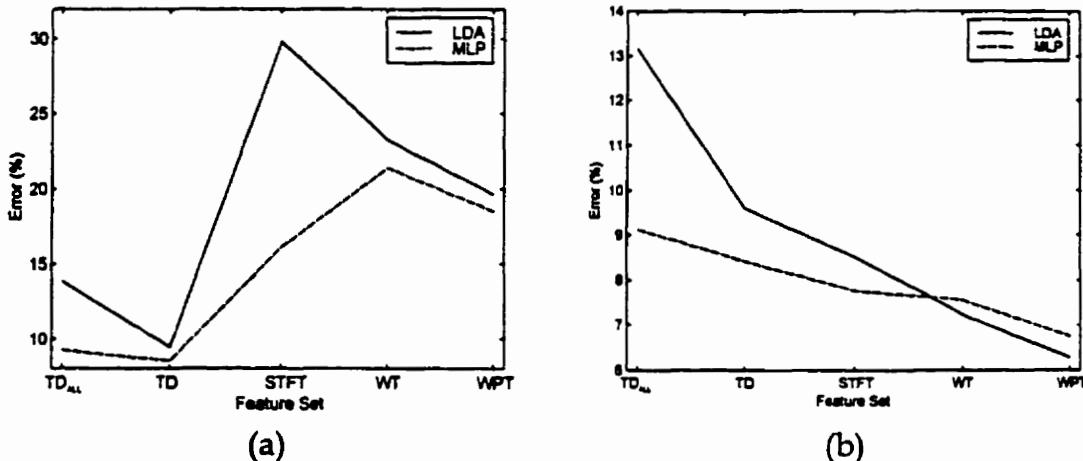


Figure 4.44 – The test set classification error, averaged across all subjects. The results are shown for each of the feature sets: (a) when using CS and (b) when using PCA. The performance when using the full time domain set (TD_{ALL}) has been included as well.

In Figure (a) it is evident that CS performs poorly when acting upon the TFR based feature sets. The transient MES has substantial dispersion in the time-frequency domain and therefore, feature selection methods fail to incorporate sufficient information for discrimination, except in very high dimensional representations. The performance is especially poor when using a LDA classifier due its inability to handle high dimensional data.

In Figure (b), when using PCA, the TFR feature sets demonstrate an obvious advantage over the TD sets. When using a LDA classifier, classification performance improves as one progresses from the unreduced TD to PCA-reduced TD features, and from STFT to WT to WPT based features. A similar trend is apparent when using a MLP classifier, but the improvement is less dramatic. This is because there are two opposing factors at work when using a MLP. To its advantage, the MLP has the ability to construct nonlinear decision surfaces. This is evident when applied to the TD feature set which, due to their lower dimension, are likely to have nonlinear characteristics. The drawback to the MLP is its

stochastic learning algorithm; an improper stopping point, a suboptimal network size, or a local minimum may deteriorate the average performance. In the progression from TD to STFT to the WT/WPT features, the dimensionality of the original feature space increases. The class boundaries in these high-dimensional spaces are likely to be essentially linear, and the PCA dimensionality reduction must preserve these linearities. In this situation, the MLP does not hold a significant advantage over the LDA. This explains why the MLP exhibits a distinct advantage when using TD features, and why this advantage diminishes when using the high-dimensional TFRs.

A properly trained MLP can always match or exceed the performance of a LDA but, since the network size and training duration was fixed for all subjects, the LDA actually outperforms the MLP when using the high-dimensional WT and WPT representations. The LDA's simplicity and good performance when using PCA-reduced features make it an attractive choice here. Overall, the best performance (6.25% error or 93.75% accuracy) is achieved when using a LDA to classify a PCA-reduced WPT feature set.

The obvious goal here is to describe the relative efficacy of each feature set as a generalized basis for classification of the transient MES. The most powerful assertion of proof would be to perform an analysis of variance (ANOVA) to prove that there is a significant difference in the mean response between the feature sets and subsequently, to perform *ad hoc* tests to prove which individual feature sets have a mean response that is significantly different than the others. Figure 4.45 shows the mean response with respect to feature set, superimposed by a scatterplot of the response from each individual subject. Only the PCA results are shown, since they represent the obviously superior method.

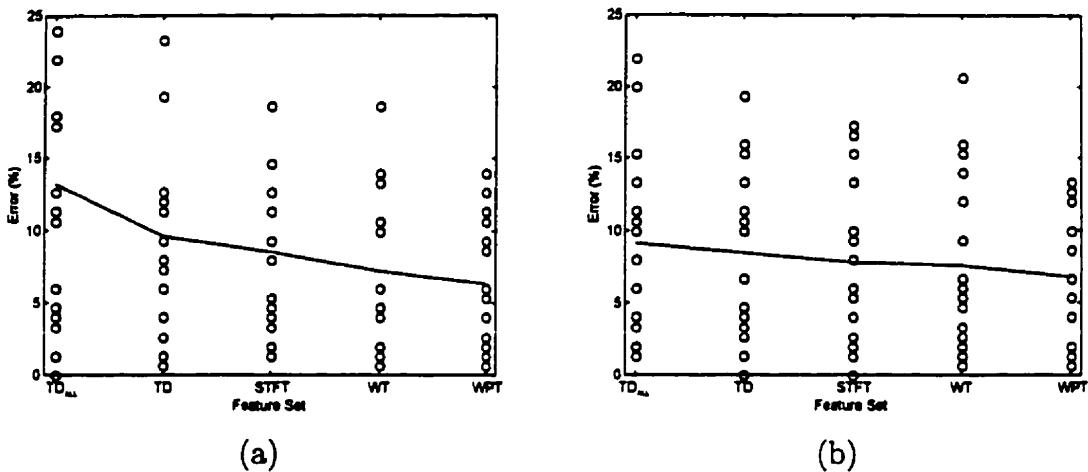


Figure 4.45 – The scatterplot of the test set classification error, superimposed by the average error across all subjects. The results are shown for PCA-reduced feature sets, (a) using a LDA classifier, and (b) using a MLP classifier. The performance when using the full time domain set (TD_{ALL}) has been included as well.

A substantial degree of variance is evident in the data. There are two sources of variance:

- 1) **Inter-subject variance.** The ability amongst subjects to reproduce patterns varies tremendously. This has an effect upon all feature sets, determining a “vertical offset” specific to the individual in the figure above.
- 2) **Inter-feature variance.** For a given subject, the relative performance of each feature set varies. This has the effect of altering the “shape” of the error versus feature set characteristic for each subject.

The hope here is that most of the variability is due to inter-subject variance, and not inter-feature variance. This would allow a significant effect due to feature set to be shown. This implies that an ANOVA must include an effect due to *Feature Set* and due to *Subject*. This model would be:

$$\text{Response} = \text{mean} + \text{Subject} + \text{Feature Set} + \text{Subject} * \text{Feature Set} + \text{error}$$

where *Subject* would be considered a random effect and *Feature Set* would be considered a fixed effect. To include the interaction term *Subject * Feature Set*, however, requires multiple instances of the parameter *Subject*. This implies multiple trials for each subject. Multiple trials were not possible however, due to the time required to acquire, and the computational resources required to process these data.

Consider for now the results when using PCA dimensionality reduction with a LDA classifier¹. If a two-way ANOVA is performed without the interaction term, treating *Feature Set* as a fixed effect and *Subject* as a random effect, it can be shown that there is a significant effect due to *Feature Set*² ($p = 0.000, r^2 = 0.125$), and even more so due to *Subject* ($p = 0.000, r^2 = 0.732$). A Scheffe *post hoc* test [Hicks93] shows significant differences ($\alpha = 0.05$) between TD_{ALL} and all other features, and between WPT and both TD_{ALL}/TD. The lack of an interaction term, however, limits the power of this test.

Another approach would be to remove the effect due to inter-subject variance by some analytic means. This can be done by normalizing the response of each subject across all feature sets. Consider, for each subject i , a normalization of the error for each reduced feature set:

$$\overline{\text{TD}}_{\text{ALL},i} = \frac{\text{TD}_{\text{ALL},i}}{\Sigma_i}, \quad \overline{\text{TD}}_i = \frac{\text{TD}_i}{\Sigma_i}, \quad \overline{\text{STFT}}_i = \frac{\text{STFT}_i}{\Sigma_i}, \quad \overline{\text{WT}}_i = \frac{\text{WT}_i}{\Sigma_i}, \quad \overline{\text{WPT}}_i = \frac{\text{WPT}_i}{\Sigma_i}, \quad (4.12)$$

¹ The CS results need not be further scrutinized, as they are clearly inferior to the PCA results.

² The r^2 value for a given effect is defined as the sum of squared errors for that effect, divided by the total sum of squared errors in the model.

where

$$\Sigma_i = TD_{ALLi} + TD_i + STFT_i + WT_i + WPT_i \quad (4.13)$$

provides a measure of the overall “ability” of subject i . This normalizes an individual’s response and conveys only the relative performance of each feature set, allowing the response from each subject to be interpreted on the same scale.

Figure 4.46 depicts the normalized test set classification error:

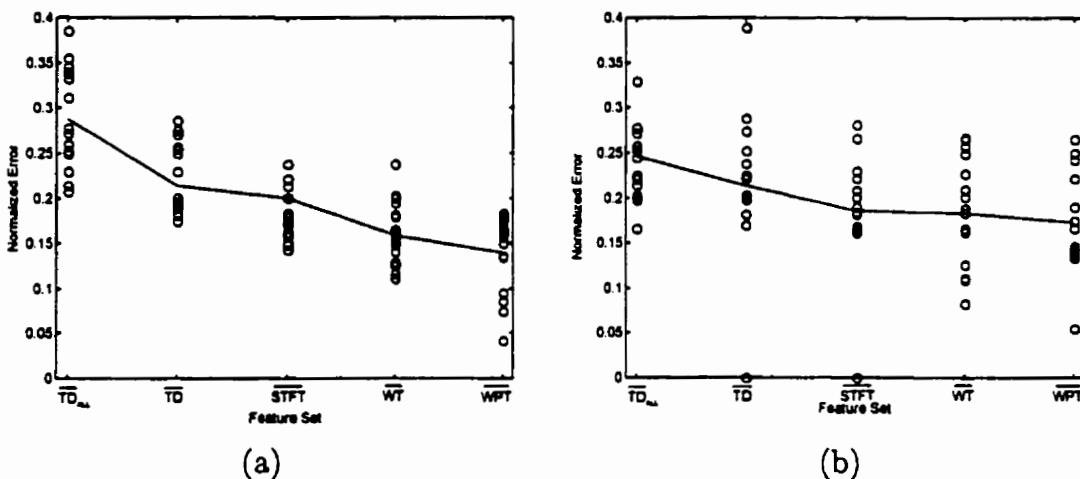


Figure 4.46 – The normalized classification error of the test set. The results are shown for (a) a LDA classifier, and (b) a MLP classifier.

After normalization, it is evident that the error for each feature set is more tightly clustered, indicating that the inter-subject variance has (at least partially) been accommodated. Indeed, when using the LDA classifier, a two-way ANOVA acting upon the normalized data (treating *Feature Set* as a fixed effect and *Subject* as a random effect) indicates that there is a significant effect due to *Feature Set* ($p = 0.000, r^2 = 0.61$), but virtually none due to *Subject* ($p = 1.000, r^2 = 0.0081$).

Having established that a substantial portion of the variance due to *Subject* has been accommodated by normalization, the analysis may be redefined in terms of a one-way ANOVA. In this case, the normalized classification error is the dependent variable and the *Feature Set* is the factor. When using a LDA classifier, one can

easily reject the null hypothesis of equal means ($p = 0.000$) and state that there is a significant effect due to feature set. A Sheffe *post hoc* test reveals the following significance levels in multiple comparisons.

| Feature Set 1 | Feature Set 2 | Significance, p |
|-------------------------------------|-------------------------------------|-------------------|
| $\overline{\text{TD}}_{\text{ALL}}$ | $\overline{\text{TD}}$ | .001 |
| | $\overline{\text{STFT}}$ | .000 |
| | $\overline{\text{WT}}$ | .000 |
| | $\overline{\text{WPT}}$ | .000 |
| $\overline{\text{TD}}$ | $\overline{\text{TD}}_{\text{ALL}}$ | .001 |
| | $\overline{\text{STFT}}$ | .372 |
| | $\overline{\text{WT}}$ | .012 |
| | $\overline{\text{WPT}}$ | .000 |
| $\overline{\text{STFT}}$ | $\overline{\text{TD}}_{\text{ALL}}$ | .000 |
| | $\overline{\text{TD}}$ | .372 |
| | $\overline{\text{WT}}$ | .608 |
| | $\overline{\text{WPT}}$ | .044 |
| $\overline{\text{WT}}$ | $\overline{\text{TD}}_{\text{ALL}}$ | .000 |
| | $\overline{\text{TD}}$ | .012 |
| | $\overline{\text{STFT}}$ | .608 |
| | $\overline{\text{WPT}}$ | .655 |
| $\overline{\text{WPT}}$ | $\overline{\text{TD}}_{\text{ALL}}$ | .000 |
| | $\overline{\text{TD}}$ | .000 |
| | $\overline{\text{STFT}}$ | .044 |
| | $\overline{\text{WT}}$ | .655 |

Table 4.3 – The Scheffe test results of the test set error, with a factor of feature set. The grayed entries represent a mean difference that is significant at the $\alpha = 0.05$ level.

This can also be expressed in terms of the sets which may be considered homogeneous (the means may not be considered significantly different):

| Feature Set | Homogeneous Subset ($\alpha = 0.05$) | | |
|-------------------------------------|--|----------|----------|
| | 1 | 2 | 3 |
| $\overline{\text{TD}}_{\text{ALL}}$ | | | |
| $\overline{\text{TD}}$ | $p=.655$ | | |
| $\overline{\text{STFT}}$ | | $p=.608$ | |
| $\overline{\text{WT}}$ | | | $p=.372$ |
| $\overline{\text{WPT}}$ | | | |

Table 4.4 – The subsets of the feature sets that may be considered homogeneous as a result is the Scheffe test upon the LDA test set error. The grayed entries represent subsets within which the null hypothesis of equal means may not be rejected at the $\alpha = 0.05$ level. The probability that the subsets are indeed homogeneous is given for each pair.

These results imply that

1. A significant improvement is gained by using PCA reduction on the TD features: $\text{TD}_{\text{ALL}} \rightarrow \text{TD}$.
2. All PCA-reduced TFR based feature sets are significantly superior to TD_{ALL} .
3. A distinct trend toward improvement is evident in the progression $\text{TD}_{\text{ALL}} \rightarrow \text{TD} \rightarrow \text{STFT} \rightarrow \text{WT} \rightarrow \text{WPT}$.
4. A stronger statement can be made about the relative performance of the features sets with respect to relative (normalized) scores than absolute scores. When using a PCA/LDA combination, the relative scores subtend no significant improvement between adjacent feature sets: $\text{TD} \rightarrow \text{STFT}$, $\text{STFT} \rightarrow \text{WT}$, and $\text{WT} \rightarrow \text{WPT}$, but a significant difference amongst all other combinations.

The same analysis was performed upon the results using the MLP. The results within each feature set are not as tightly clustered as those of the LDA, due to the stochastic nature of the MLP's learning algorithm. Applying a one-way ANOVA

to the normalized scores yields $p = 0.029$, allowing one to reject the null hypothesis at $\alpha = 0.05$ across all feature sets. A Sheffe *post hoc* test however, indicates no significant differences amongst any of the individual feature sets.

Another expression of the relative utility of the feature sets is the probability that they will yield the best (or worst) performance. For each subject, the best and worst feature set was determined from the absolute classification error scores (in the case of a tie, multiple feature sets were registered). Figure 4.47 shows a histogram of the occurrence of each feature set as the best and worst feature set. Only the PCA-reduced features have been included in this analysis³.

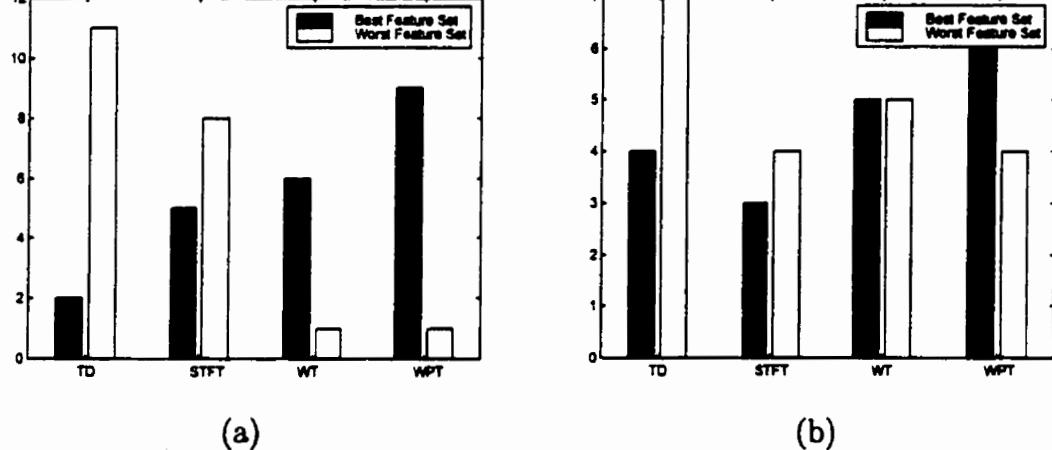


Figure 4.47 – A histogram of the occurrence of each feature set yielding the best and worst performance of all sets, accumulated over all 16 subjects. Figure (a) LDA classifier, (b) MLP classifier.

These histograms may be interpreted as sample probability density estimates. From Figure (a), it is clear that when using a LDA classifier, the likelihood of being the best feature set increases and the likelihood of being the worst feature set decreases, as one progresses from TD to STFT to WT to WPT. In Figure (b), the

³ An unreduced TD set would have dominated the measure of worst feature set, obscuring the relative performance of the others.

results are not as clear when using a MLP classifier, but it is evident that the TFR based feature sets are more likely to be the best feature set as the TD features, and are less likely to be the worst. The WPT has the highest probability of being the best feature set, and the lowest probability of being the worst.

This probability-based measure of feature set efficacy does not convey a measure of performance, but it does assign a degree of confidence to each feature set. This is a useful complement to the absolute performance given previously. These results strongly suggest the value of TFR feature sets, especially when using a LDA classifier. Once again, the combination yielding the best generalized performance is that of WPT/PCA/LDA.

4.6.2 Summary

It has been shown that, when using TFR based feature sets, PCA provides a far more effective means of dimensionality reduction than feature selection by CS. Moreover, by preprocessing the feature set with PCA prior to classification, a LDA – a classifier that is easier to implement and to train than a MLP – may be used without degrading performance. It has also been demonstrated that when using a PCA/LDA combination, there is a significant improvement in performance in the progression $\text{TD}_{\text{ALL}} \rightarrow \text{TD} \rightarrow \text{STFT} \rightarrow \text{WT} \rightarrow \text{WPT}$. The best performance is exhibited when using a WPT/PCA/LDA combination, yielding an average classification error of 6.25%. This represents a significant improvement over Hudgins' method which, for these data, subtend an average error of 9.25%.

The improved accuracy provided by these methods will enhance the functionality of a pattern recognition based myoelectric control system. The WPT has a complexity on the order of $N \log N$, and the stages of PCA and LDA, once trained, require a simple matrix multiplication in the feedforward path. Therefore, the combination WPT/PCA/LDA easily lends itself to real-time implementation on a DSP microprocessor of modest capabilities.

Chapter 5

Assessment of Classification Performance

The results of Chapter 4 have provided some interesting observations about the transient MES classification problem. Projection-based dimensionality reduction (in the form of PCA) has been shown to be markedly superior to feature selection methods. As well, an improvement in performance due to the feature set has been shown in the progression TD \rightarrow STFT \rightarrow WT \rightarrow WPT. The purpose of this chapter is to provide some perspective on these results.

In Section 5.1: *Generalized Feature Dimension*, it will be shown that the process of estimating feature set dimension by validation can be avoided without sacrificing performance. This reduces the time required to acquire and process the data. In Section 5.2: *Performance Bounds*, evidence is accumulated in an effort to show that the average performance of the WPT/PCA representation may be approaching the Bayes error. Section 5.3: *Modeling the Transient MES Classification Problem* investigates several models of the intra-class variance that characterizes the transient MES classification problem. A model based upon the WT is shown to emulate the structural variance of the transient waveform. The

appropriateness of this model is confirmed by the fact that it places demands upon the signal representation that are similar to the real MES datasets. Further, the simulated data is shown to effectively augment sparse training sets, improving generalization performance.

5.1 Generalized Feature Dimension

Thus far, the premise of selecting the feature set dimension has been on an *individual basis*; specifying the dimension which minimizes the classification error of a validation set. An important question is: how crucial is this validation process? Certainly, validation comes at the expense of significant computational effort (determining the classification error of the validation set at all possible feature set dimensions) and additional data collection. Is it possible that some *generalized feature set dimension* – a fixed dimensionality for all subjects – may serve just as well? A generalized feature set dimension could be that which minimizes the ensemble averaged classification error across the subject database.

An illustration of these two dimensionality specification schemes is shown in Figure 5.1. Consider, for example, a feature set of STFT coefficients reduced using class separability. The generalized dimension is chosen as that which minimizes the validation set error, averaged across all subjects.

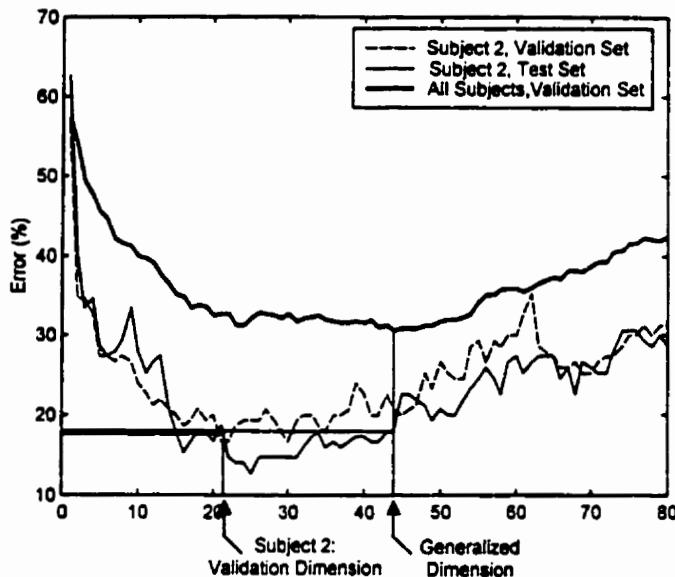


Figure 5.1 – Dimensionality specification on an individual basis versus that using a generalized dimension. The test set error of Subject 2 is shown when using the “optimal” dimension chosen from the validation set of Subject 2, and when using the generalized dimension, determined from the averaged response of all subjects.

In this example (using data from Subject 2), there is little difference between the test set error evaluated at the individually specified dimension and the generalized dimension.

It is instructive at this point to recall the nature of the relationship between classification error and feature set dimension. This relationship depends upon not only the subject, but upon the feature set, the means of dimensionality reduction, and the classifier. The validation set classification error was computed at all possible dimensions (up to a maximum of 64) for each subject, using each feature set. Furthermore, the performance was determined using both CS and PCA dimensionality reduction, and using both LDA and MLP classifiers. Figure 5.2 depicts the overlay of the validation set classification error for each subject using all possible combinations.

In Figure (a), CS dimensionality reduction and a LDA classifier are used. As previously described, the TD feature set easily outperforms the TFR feature sets.

As is evident in Figure (a), the performance of the LDA degrades as the dimension grows larger than 40. Obviously, the combination of CS and LDA is inappropriate when using TFR feature sets.

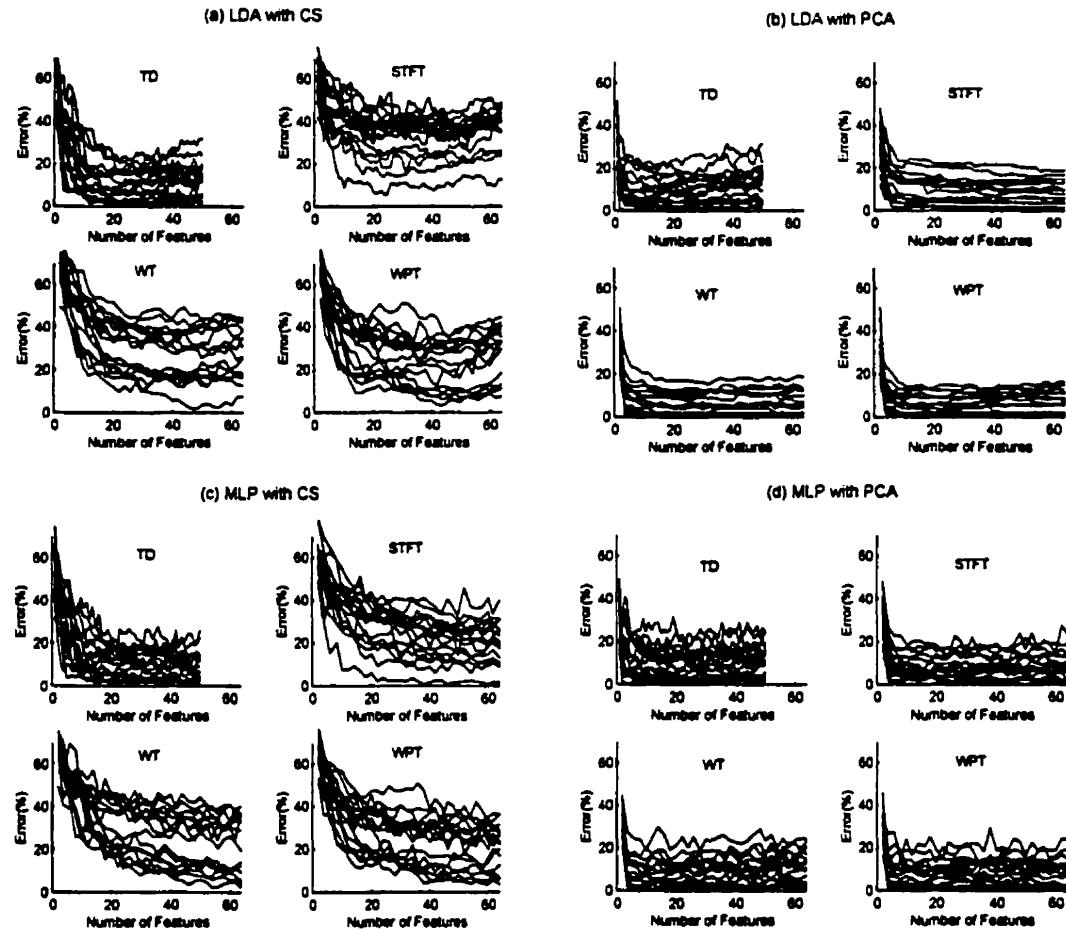


Figure 5.2 – The effect of feature set dimension upon validation set classification error, for all subjects in the two channel database. For each of Figures (a)-(d), the error for each subject is shown for each feature set. The results are shown in the context of the dimensionality reduction method and the classifier: (a) CS dimensionality reduction is used with a LDA classifier; (b) PCA and LDA; (c) CS and MLP; (d) PCA and MLP.

In Figure (b), PCA dimensionality reduction and a LDA classifier subtend much better results when using the TFR feature sets. Moreover, the relationship between classification error and feature set dimension is much more regulated. The error sharply declines until, at a dimension of about 20, the response flattens and does not experience nearly the variance with increasing dimension that is

evident in Figure (a). This implies that classification performance is much less sensitive to dimensionality specification when using PCA dimensionality reduction.

In Figure (c), CS dimensionality reduction and a MLP classifier yields results that are somewhat better than that when using CS with a LDA (Figure(a)), since a MLP can accommodate a higher dimension input than a LDA. This is evident in that the error continues to decline with increasing dimension of the TFR sets¹. The performance, however, does not match that of PCA with LDA (Figure (b)).

In Figure (d), the MLP is used with PCA-reduced feature sets. The performance is similar to that in Figure (b) when using a LDA classifier, except that the variability with respect to feature set dimension is greater. This is due to the stochastic nature of backpropagation learning; the MLP training algorithm introduces additional variance into the solution.

For each subject, the feature set dimension must be determined as that which minimizes the validation set dimension. Figure 5.3 shows the chosen dimensionality for each feature set, when using each combination of dimensionality reduction method and classifier.

¹ This error rate asymptotically flattens at a dimension of about 64; further declination is negligible.

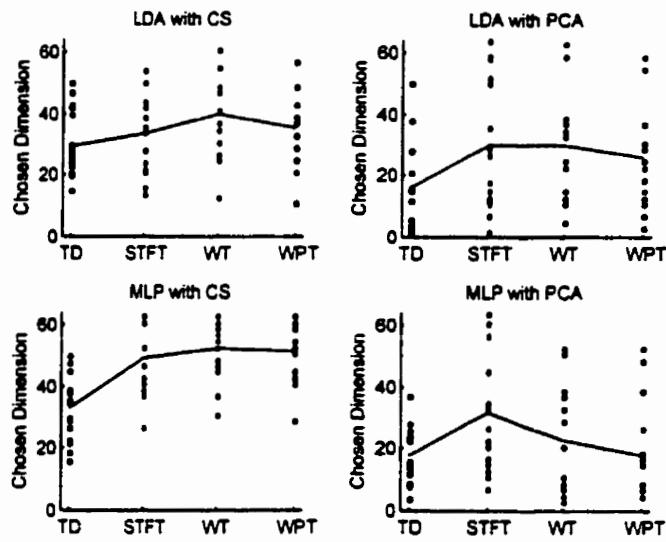


Figure 5.3 – The chosen dimension when using each combination of feature set, dimensionality reduction method, and classifier. The points represent the results of individual subjects; the solid lines represent the mean across all subjects.

The PCA-reduced feature sets exhibit a greater degree of variance in the chosen dimension. This is due to the relatively regulated association between classification error and feature set dimension. On average, CS demands a greater feature set dimension than PCA.

The interesting question is, however, how does a generalized dimension compare to validation based dimensionality specification? The generalized dimension can be based upon empirical evidence; the most logical measure would be the dimension which minimizes the validation set classification error, averaged across all subjects. Figure 5.4 depicts this averaged response for each combination of feature set, dimensionality reduction method, and classifier.

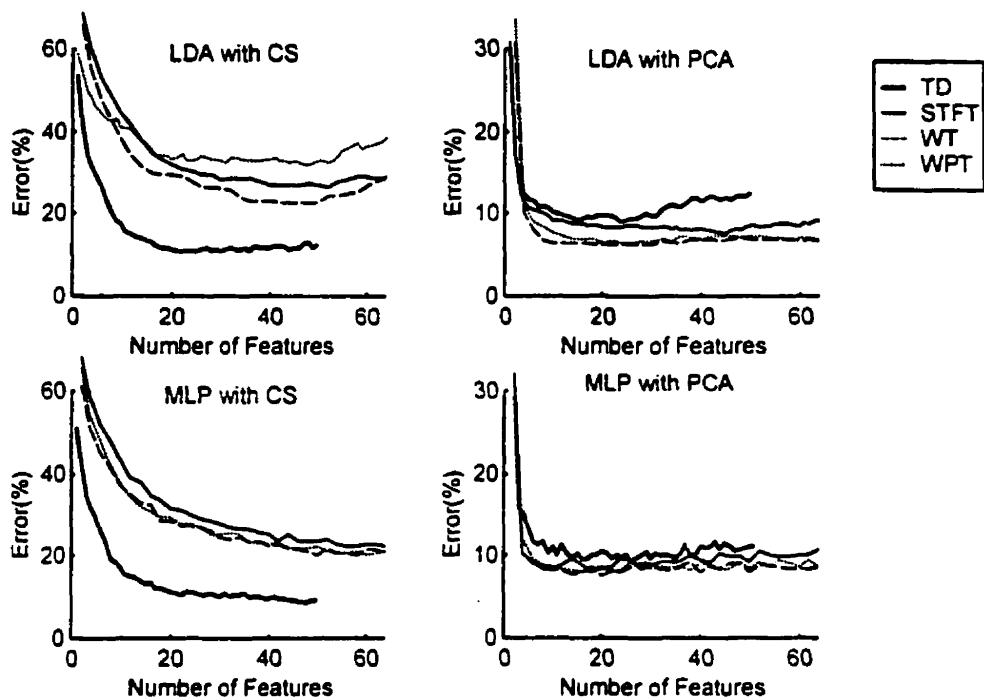


Figure 5.4 – The effect of feature set dimension upon the validation set error, averaged across all subjects.

From the information in this plot, one may specify a generalized dimension as that which minimizes the averaged validation set error for each dimensionality reduction/classifier combination. Rather than selecting the exact minima, the generalized dimension will be specified as a coarse estimate, rounded to the nearest decade, so as to avoid the implication that the generalized dimension must be a carefully measured quantity:

| Method | Generalized Dimension | |
|--------------|-----------------------|---------|
| LDA with CS | TD: 30 | TFR: 40 |
| LDA with PCA | TD: 20 | TFR: 30 |
| MLP with CS | TD: 50 | TFR: 60 |
| MLP with PCA | TD: 20 | TFR: 20 |

These generalized dimensions were used as the test set dimension; the resulting classification error was compared to that when using the individually specified

(validation-based) dimensions. A comparison of these two approaches to dimensionality specification is shown Figure 5.5.

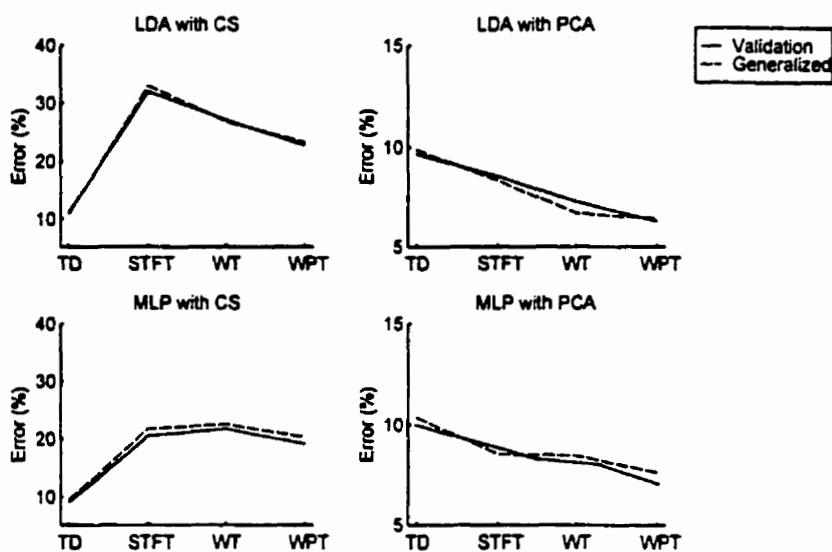


Figure 5.5 – A comparison of the test set classification error when using individually specified (validation-based) dimensionality and generalized dimensionality. The averaged error across all subjects is shown for each feature set and dimensionality reduction/classifier combination.

Clearly, a generalized dimension subtends an error rate that is very similar to that when using individually specified dimensions. There is clearly no advantage to specifying dimensionality for each subject on an individual basis, regardless of the classifier or dimensionality reduction method. Obviously, an individual specification of dimensionality would always do as well or better than a generalized dimensionality specification if the validation set error exactly matched that of the test set. Rather, it is merely an *estimate* of the test set's dependence on dimensionality, subject to statistical variance due to a finite sample size (a limited number of patterns within the validation set and the test set). A larger number of patterns in each set would decrease the variance of this estimate, but it is unreasonable to expect that one could acquire more than that which has been used here (150 in each set) in a clinical situation.

In order to determine the generalized dimensionality, the existence of a validation dataset for each subject has been assumed, and the classification error has been computed at each dimension, for each subject. Although this is an involved process, it does not have to be repeated when analyzing new subject data. The implicit assumption here is that the averaged response of this subject database generalizes well to each subject within the database, and to novel datasets of two channel transient MES. Therefore, having chosen a generalized dimension for a particular feature set/dimensionality reduction method/classifier combination, no validation dataset and no validation process is necessary when acquiring new data in a clinical situation.

5.2 Performance Bounds

The results of the previous chapter indicate the relative performance of various feature sets, dimensionality reduction schemes and classifiers. From this, the best strategies for classification of the transient MES were determined. The most meaningful assessment of these methods however must address the question: *what is the best performance that one can expect of this problem?* If we limit the range of possibilities to statistical signal representation then the answer is: *we can do no better than a Bayes classifier.* By its definition, the Bayes classifier is that which yields the lowest probability of error.

For a given subject, this lower bound is determined by the separability of the data: specifically, the degree of overlap in the class-conditional density functions $p(\mathbf{x}|y_k)$ (refer to Section 2.3.1). This separability (or complexity), in turn, depends upon

- 1) *Operator ability.* This is the skill with which a subject can consistently reproduce an intended contraction pattern.
- 2) *Operator error.* Somewhat different than operator ability, operator error is defined to be any improper or inadvertent contractions that the subject may have produced.
- 3) *System error.* These are factors due to the instrumentation, which affect the quality of the recording. These include electrode movement, inductive spikes due to lead motion, instrumentation noise, and environmental noise. These have the effect of improperly triggering the system or introducing an artifact into the recording.
- 4) *The proper selection of a threshold trigger.* Is the most informative portion of the waveform captured for analysis?

The ultimate evaluation of the performance of the methods prescribed here would be in comparison to the Bayes error. This comparison may be performed in the following ways (in descending order of robustness):

1. *Determine an analytical form of the Bayes error.* This requires an analytical expression of the class-conditional probabilities. This is not possible for most physical datasets due the absence of a complete mathematical model.
2. *Estimate the Bayes error.* Although this is possible for datasets of low dimension, the class-conditional densities are extremely difficult to estimate for dimensions higher than ten [Scott92]. Any density estimate upon the 256-dimensional transient MES data would require an inordinately large number of exemplars to be stable.
3. *Provide evidence that the average error cannot be reduced.* Consider a scheme of feature extraction/dimensionality reduction/classification that is to be scrutinized. If a number of strategies that suggest an improvement do not have an effect on classification error, then an implication may be made that the performance is at or near the Bayes limit.

Methods (1) and (2) are simply not feasible for the transient MES classification problem. The approach of method (3) has been taken here to qualify the performance of the WPT/PCA signal representation. The following sections describe augmentative strategies which seek to improve upon the results of Chapter 4.

5.2.1 Hybrid Feature Sets

Although wavelet packet based features have demonstrated the best overall performance, they are not the best for each individual. Also, even if a WPT is the best feature set for a given subject, there may be information in the other feature sets that is exclusive of the WPT representation. This is the motivation for combining the features to form a hybrid feature set.

A hybrid feature set was constructed in the following manner. The TD, STFT WT and WPT features were computed, and a hybrid feature set was compiled using either CS or PCA dimensionality reduction. The reduced features were then ranked, and interleaved so that the most important features (within each feature set) are ordered first ($TD_1, STFT_1, WT_1, WPT_1$), followed by the second most informative features ($TD_2, STFT_2, WT_2, WPT_2$), and so on, to form an aggregate feature set. This is illustrated in Figure 5.6.

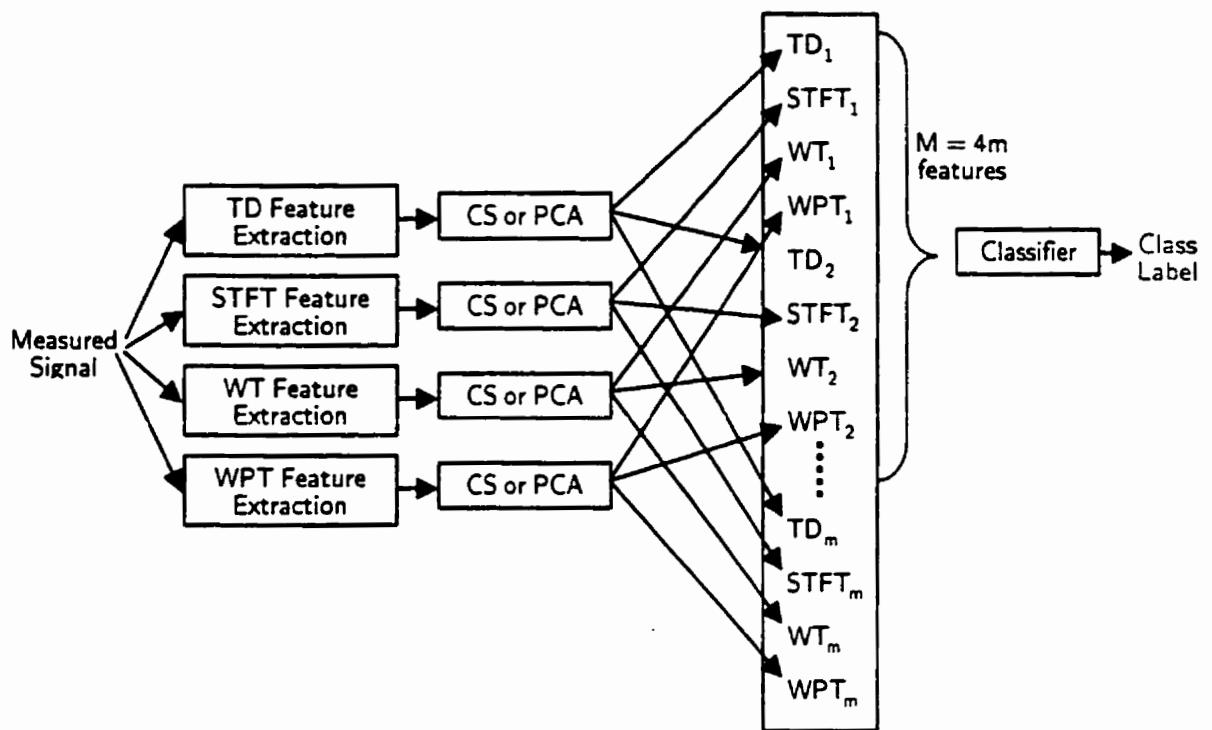


Figure 5.6 – A hybrid feature set. Each feature set is subject to dimensionality reduction, and the reduced features are ranked (denoted by the subscripts $1, 2, \dots, m$). These ranked features are interleaved to form the aggregate feature set, from which M features are used in the representation.

As before, the dimension is selected as that which minimizes the validation set error. Figure 5.7 shows the test set classification error for the hybrid feature set as compared to the individual feature sets, averaged across all subjects.

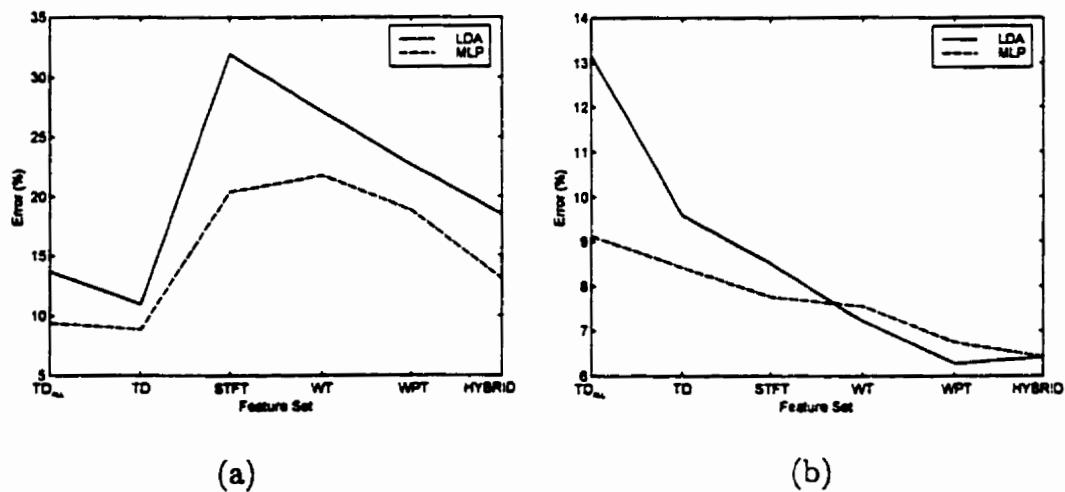


Figure 5.7 – The test set classification error, averaged across all subjects, when using a hybrid feature set. Figure (a) shows the results when using CS, and (b) when using PCA.

In Figure (a) when using CS, the hybrid set performs better than any of the TFR sets, but worse than the TD set. This is because the TFR coefficients are extremely noisy, and they distract the classifier from the otherwise stable TD coefficients.

Figure (b) shows the results when using PCA. Interestingly, the hybrid set has very nearly the same performance as the WPT feature set. Although a slight improvement is exhibited by the MLP, the performance of the LDA decreases. One might expect that a hybrid set would outperform the WPT set, if the hybrid set contributes original information.

Another augmentation of the PCA-reduced feature set suggests a further improvement in the efficacy of representation. The interleaved features are ordered with respect to their importance within each feature set, but their overall order is undetermined (there is no guarantee that STFT₃ is more informative than TD₄, for example). Also, the PCA-reduced hybrid feature set most likely contains redundancies amongst the individually projected TD, STFT, WT and WPT features. Unneeded coefficients may degrade classifier generalization if added dimensionality contributes no novel information. The PCA-reduced hybrid set (at all possible dimensions) was therefore subject to a second stage of PCA to remove linear dependencies. This approach however, yields essentially the same results as the original PCA-reduced hybrid set.

The fact that a PCA-reduced hybrid feature set can match but not surpass the performance of a WPT feature set suggests that a negligible amount of new information is introduced into the signal representation by the TD, STFT and WT

features. A second stage of PCA, intended to remove unnecessary linear dependencies in the hybrid set, has no effect. The strongest implication of these two observations is that the PCA-reduced WPT representation may be yielding results that are near the average Bayes error for the ensemble of subjects.

5.2.2 Two-Class LDB Formulation

The LDB algorithm has improved the WPT representation for classification by providing an orthonormal basis determined by a class separability measure, rather than a criterion based upon signal reconstruction. This section suggests an improvement to the LDB algorithm, and shows the effect upon the transient MES dataset.

In an attempt to improve the effectiveness of the LDB algorithm, the 4-class problem was broken down into an ensemble of four 2-class problems. If the classes are identified as A, B, C and D , four LDBs are constructed which correspond to:

$$\begin{aligned} \text{LDB A: class } A &\text{ or } \{B, C, D\} \\ \text{LDB B: class } B &\text{ or } \{A, C, D\} \\ \text{LDB C: class } C &\text{ or } \{A, B, D\} \\ \text{LDB D: class } D &\text{ or } \{A, B, C\} \end{aligned} \quad (5.1)$$

The motivation for constructing an ensemble of 2-class problems is that the measure of class separability used to prune the wavelet packet tree becomes obscured as the number of classes increases. In this scheme, the resulting 2-class LDBs will provide a more appropriate tiling for each 2-class problem. Each LDB is specifically tuned to isolate a particular class from the others. Therefore, when a

pattern from an unknown class is presented to the collection of LDBs, there will be one LDB that is likely to maximally separate the pattern from the others. It is anticipated that the existence of a single well-tuned LDB in a set of poorly-tuned LDBs will be more informative than that of a single 4-class LDB.

The WPT coefficients from each LDB were subject to a dimensionality reduction stage (CS or PCA), and presented to a classifier (LDA or MLP). For LDB *A*, the classifier returns two outputs that are analogous to the probability that the pattern *belongs to class A* or *does not belong to class A*². The class is assigned to the classifier that has the largest response to the class it isolates. This scheme is depicted in Figure 5.8.

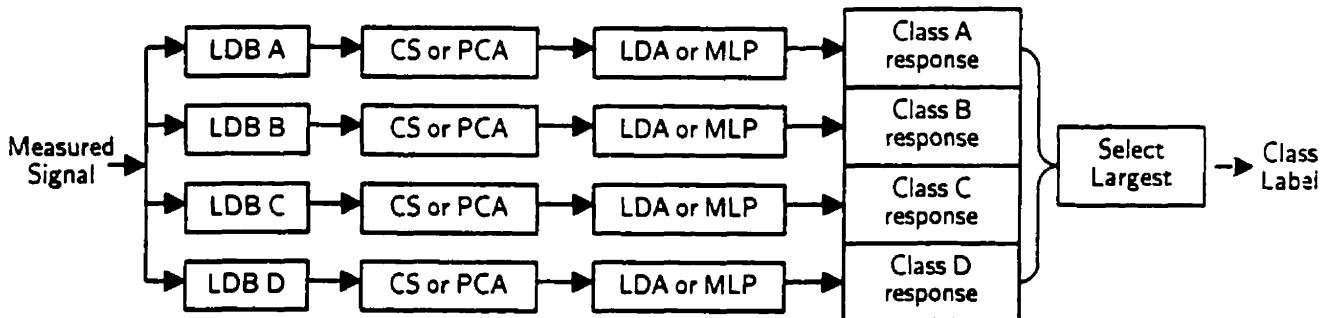


Figure 5.8 – A classification scheme implementing an ensemble of 2-class LDBs.

The results when using this approach are compared to those when using the original 4-class LDB in Figure 5.9.

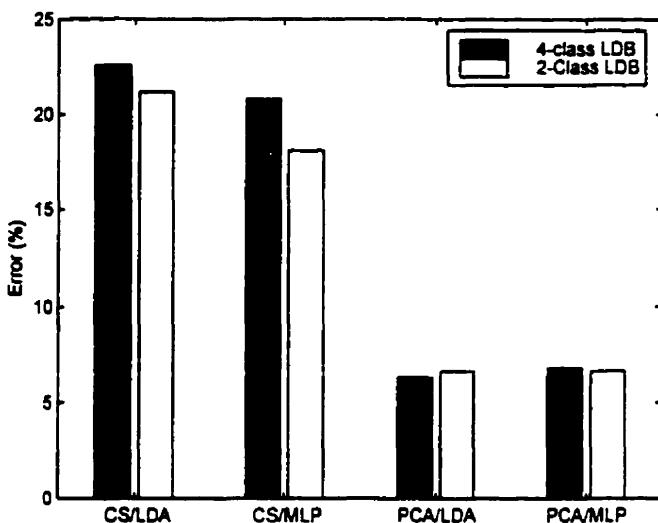


Figure 5.9 – The relative performance of 4-class and 2-class LDB schemes, averaged across all subjects. The results are shown for each combination of CS/PCA dimensionality reduction and LDA/MLP classifiers.

The 2-class LDB configuration demonstrates an improvement over the 4-class LDB when using CS dimensionality reduction. This is true for both the LDA and the MLP classifiers. The 2-class configuration appears to concentrate the information important for discrimination more effectively. When using PCA-reduced features however, the 2-class LDB configuration does not seem to have much effect at all. The PCA dimensionality reduction scheme appears capable of consolidating the information in a 4-class LDB just as effectively as in an ensemble of 2-class LDBs.

Once again, a proposed refinement of the signal representation yields improved performance when using CS, but not when using PCA. This is another implication that the PCA/WPT representation is performing at a near-optimal level.

² The MLP, by its construction, yields an output that may be interpreted in this manner. To have the LDAs

5.2.3 Bootstrapping with Noise

The bootstrap technique is a tool for augmenting training data by sampling with replacement. A variant of the simple bootstrap involves the injection of noise into the data when sampling from the original training set. This procedure – *bootstrapping with noise* or *training with jitter* – has been shown to improve the generalization properties of an estimator [Sietsma91]. A simplistic view of bootstrapping is that it attempts to simulate the noise inherent in the data, and therefore, effectively increase the number of training patterns. A more precise characterization is that using noisy replications of the training data is closely related to kernel regression estimation, such that it regulates the tendency of an estimator to overfit the data [Scott92]. It has been shown that training with small amounts of noise controls the smoothness of a classifier’s mapping function, amounting to a form of smoothness regularization [Bishop95]. Bootstrapping with noise works for classification because the functions (the decision surfaces) that are desired are usually smooth surfaces.

Bootstrapping with noise was used to augment the two channel MES data. The training set was randomly sampled with replacement, and zero-mean Gaussian noise was added to each pattern. The amplitude of the noise must be large enough to have an effect, but not so large as to distract the classifier. By trial and error, the noise standard deviation was specified to be one-tenth that of the average MES

respond in the same way, the outputs were normalized to sum to unity.

pattern. Using this scheme, the training set was augmented by noisy patterns until it was twice its original size³.

Figure 5.10 shows the effects of bootstrapping with noise upon the average test set classification error when using CS-reduced feature sets.

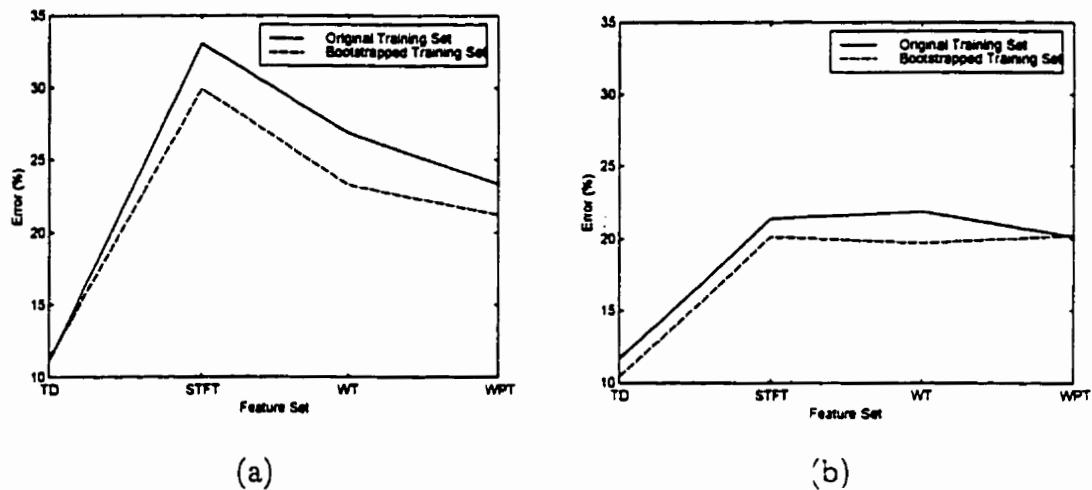


Figure 5.10 – The effect of bootstrapping with noise upon the average test set classification error when using CS dimensionality reduction. Figure (a) shows the results when using a LDA, and (b) when using a MLP.

The bootstrap technique improves the generalization performance for both classifiers, and for all feature sets (except for two cases where there is no effect). The CS-reduced features have a substantial degree of variance and as a result, the classifier will tend to overfit these data by imposing complex boundaries. Training with jitter has a smoothing effect that simplifies these boundaries, and improves generalization.

³ Certainly, the bootstrap technique can be used to produce training sets of arbitrary size. The training set size was merely doubled here (to limit processing requirements), but this is enough to demonstrate its effects upon classification performance.

Bootstrapping does not have the same effect however, when using PCA-reduced features, as shown in Figure 5.11.

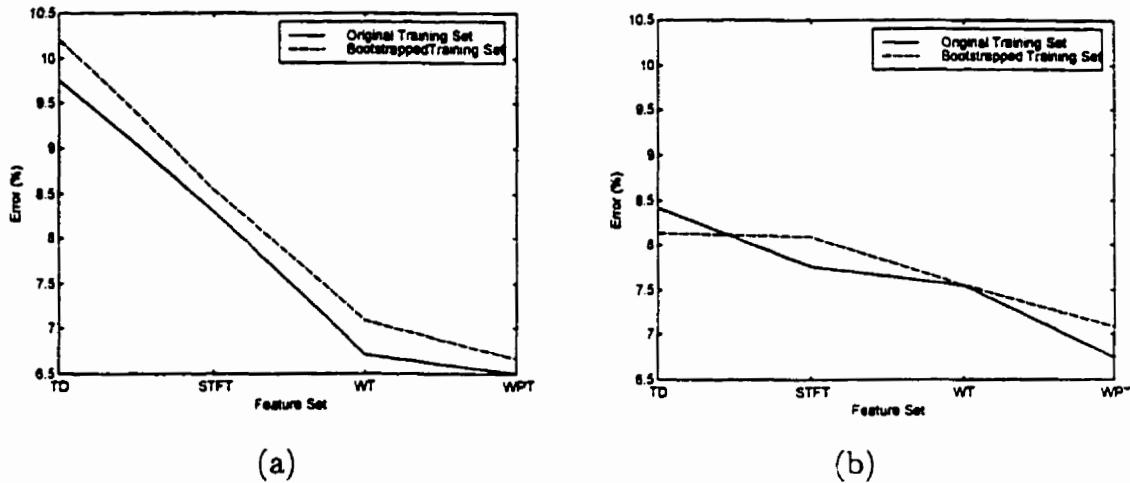


Figure 5.11 – The effect of bootstrapping with noise upon the average test set classification error when using PCA dimensionality reduction. Figure (a) shows the results when using a LDA, and (b) when using a MLP.

The bootstrap technique either slightly degrades performance, or has negligible effect. PCA tends to relegate uncorrelated noise to the lesser principal components, reducing the effect it may have on the leading components. It has been observed however that the LDA classifier, which is capable of only linear maps, usually performs as well as the MLP classifier when using PCA. This implies that the decision boundaries in the PCA-reduced feature space are simple, such that the classifiers probably need no regularization.

5.2.4 Spin Cycle Training

It has been suggested that the lack of translation invariance may degrade the performance of WT and WPT based feature sets, as shifts in the input signal will produce modifications of the transform coefficients that are much more complex

than a simple shift. This effect is nonlinear, due to the decimation-by-two operation that occurs at each stage of the decomposition (for details, see Appendix D). A method termed “cycle spinning” was proposed by Coifman *et al.* [Coifman95] to describe the technique of creating several shifted versions of the data to “average out” translation dependence when denoising signals.

The concept of cycle spinning may be applied to a classification problem as well. It will be assumed that there is some degree of temporal translation present in the training and test data for an arbitrary subject. This will introduce a nonlinear modification of the WT/WPT coefficients in both sets. Now, consider augmenting the training set by shifted versions of the original training set. The classifier is given the task of accommodating the nonlinear modification of the WT coefficients, having been given a larger set of exemplars upon which to train. This approach attempts to “show” the classifier a richer set of dispersive effects, hopefully, with the result of improved generalization.

The WT and the WPT were applied to the two channel MES data, and reduced feature sets were determined using both CS and PCA dimensionality reduction. The data from each individual were subject to spin cycling: the original training set was replicated by shifted versions of itself. A training set augmented by M spin cycles comprises the original set plus replicates shifted by $1, 2, \dots, M - 1$. Figure 5.12 shows the effects of spin cycling upon the CS and PCA reduced feature sets.

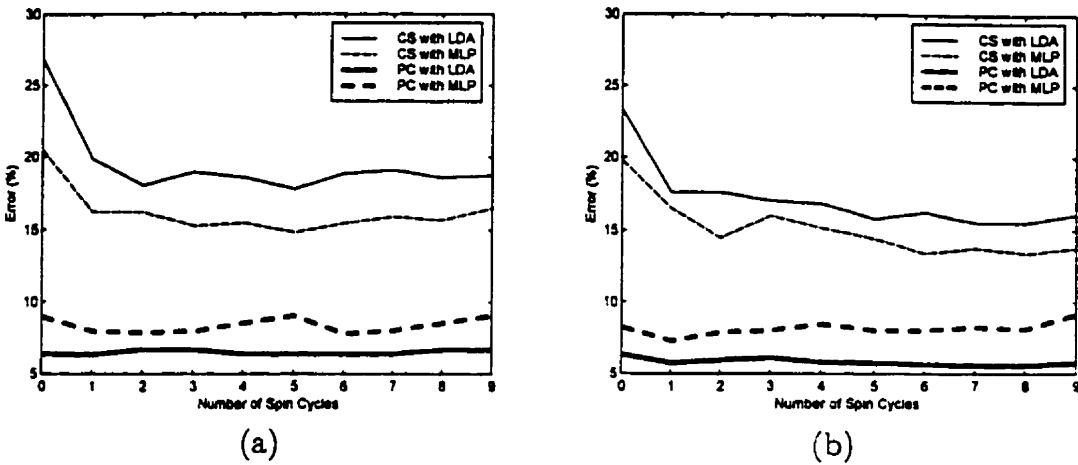


Figure 5.12 – The effects of spin cycling upon the test set classification error, averaged across all subjects. A training set augmented by M spin cycles comprises the original set plus replicates shifted by 1 to M . Here, M ranges from zero to nine. Figure (a) shows the response of CS and PCA-reduced WT features, and (b) CS and PCA-reduced WPT features.

Similar traits are evident for both the WT and the WPT. When using CS dimensionality reduction, an improvement in classification error is evident when applying a few spin cycles. Beyond five spin cycles, no obvious benefit is apparent. This is true for both the LDA and the MLP classifiers. When using PCA dimensionality reduction the performance is markedly superior to CS, but no benefit is gained by spin cycling.

An explanation of the effects of spin cycling requires an examination how shift manifests itself in the WT and WPT feature sets. If a signal is translated by a series of shifts (for example, shifts from $-10, \dots, -1, 0, 1, \dots, 10$) there is no obvious pattern of modification in the WT/WPT coefficients. Indeed, the effect seems to be that of random noise. Therefore, if the training set is augmented by shifted exemplars, this amounts to bootstrapping with noise, which was demonstrated in the previous section. This has the effect of regularizing the CS-based representation, improving its performance.

If the “noise” is uncorrelated however, it will be suppressed in the leading PCA features. Since PCA is capable of consigning this “noise” in the training set to the lesser principal components, the hope is that the projections formulated from the training set will suppress the “noise” in the test set as well. This is likely, due to the fact that any “noise” in the test set is derived from the same intrinsic shift as the training set. In this case, PCA can accommodate (at least partially) the effects of shift in the test set.

Note: In Appendix D, the “noise-like” behavior of the WT modifications due to shift is investigated. It is shown that PCA does indeed tend to remove the “noise” components from the leading PCA features, and that it does so consistently in the test set as well as the training set. Moreover, it is shown that this mechanism allows the classification performance of the PCA representation to be relatively insensitive to shift.

5.2.5 Summary

The methods in this section have been proposed to improve the classification performance of the transient MES.

To summarize these results:

A hybrid feature set can match but not exceed the performance of the WPT/PCA representation. These results suggest that it may not be possible to achieve a

lower average classification error using any combination of the feature sets (TD, STFT, WT, WPT) and dimensionality reduction methods (CS, PCA) investigated here.

A 2-class LDB configuration improves the performance when using CS-based features, but not when using PCA. This implies that a refined 2-class LDB tiling provides better concentration of discriminant information. This is beneficial when using a subset selection scheme, such as CS. PCA dimensionality reduction however, is capable of consolidating the information in a 4-class LDB just as effectively as in an ensemble of 2-class LDBs.

Bootstrapping with noise improves the performance when using CS-based features, but not when using PCA. The main effect of bootstrapping with noise is to regularize the problem, and prevent overfitting. This was shown to improve the performance of the CS-reduced feature sets. PCA has the effect of segregating the signal and noise subspaces however, eliminating the effect of bootstrapping. As well, the decision surfaces when using PCA are simple, and would not benefit from regularization.

Spin Cycling improves the performance when using CS-based features, but not when using PCA. The effect of shift manifests itself as noise in the WT and WPT coefficients. The improvement exhibited when using CS-reduced WT and WPT features is due to smoothness regularization. PCA tends to relegate these “noisy” effects of shift to the lesser principal components, and is therefore not affected by spin cycling. This implies that PCA is accommodating the effects of shift intrinsic to the training and test sets.

The implication of these observations is that the PCA-reduced WPT representation may be yielding the best results possible of a representation derived from TD, STFT, WT or WPT features. The fact that no method can improve upon the performance of the WPT/PCA combination, even though each augmentation attempts to refine the representation in a different way, suggests that the Bayes bound may be near. The fact that a MLP can do no better than a LDA suggests that the complexity of the problem has been accommodated by the signal representation and not the classifier.

As previously mentioned, embedded in the Bayes error for each subject is the operator error and system error. These errors generate extraneous patterns which cannot be accommodated by the classifiers. Patterns that were improperly triggered or were due to inadvertent activity were rejected, but no attempt was made to reject data that simply "looked" bad. Given the length of the data acquisition procedure (each subject produced 400 contractions, which required about an hour), it is certainly possible that some operator/system error may be present. Indeed, it is highly likely that these errors represent a significant portion of the observable error for any given subject.

It is possible that some other form of statistical signal representation may outperform the WPT/PCA, but it is unlikely that the difference would be substantial. The only possibility that the average error might be reduced beyond the Bayes error lies in an extremely effective form of syntactic signal representation. It is not clear if such a syntactic model is feasible.

5.3 Modeling the Transient MES Classification Problem

The demands of the signal representation described in this work are due to the nature of the within-class (*intra-class*) variance inherent in the transient MES classification problem. In this section, some models of intra-class variance are investigated in an effort to explain the behavior observed of real MES datasets.

5.3.1 Additive Noise

When analyzing the classification performance of a given application, it is common to investigate the performance of the system in the environment of additive noise. Although this is a reasonable model of intra-class variance in some applications, it is simply not the case in the transient MES classification problem. Nonetheless, an analysis of this sort provides some insight into the properties of the feature set and dimensionality reduction techniques.

A set of four patterns (one from each class) were selected from an arbitrary subject to serve as template signals. Zero-mean white Gaussian noise was added to these template signals, and replicates were formed to create training, validation and test sets equal in size to those of the real MES datasets (100, 150, and 150, respectively). The test set classification error was determined when using TD, STFT, WT and WPT features, and when using CS and PCA dimensionality reduction. Figure 5.13 shows the performance versus signal to noise ratio (*SNR*), where

$$SNR \doteq \frac{\text{Signal Power}}{\text{Noise Power}} \quad (5.2)$$

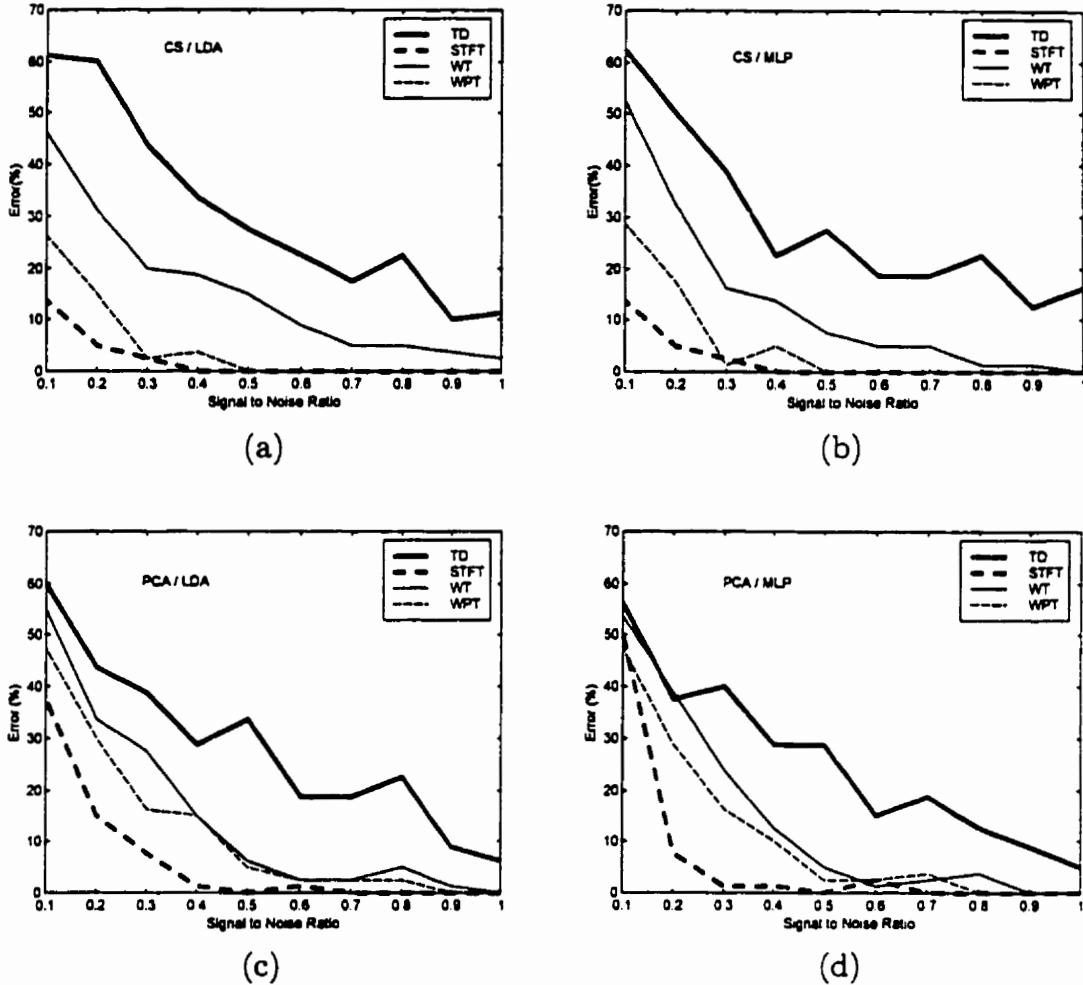


Figure 5.13 – The performance in additive noise. The test set error rate is shown for each feature set as a function of SNR. The results include (a) CS using a LDA, (b) CS using a MLP, (c) PCA using a LDA, and (d) PCA using a MLP.

The response is very similar in each plot, regardless of the form of dimensionality reduction or the classifier. The TFR feature sets do show a slight advantage when using CS instead of PCA. The STFT exhibits the best noise immunity of all feature sets, allowing roughly 88% accuracy when the noise power is ten times that of the signal when using CS, as shown in Figures (a) and (b). Because the noise is stationary and white, it is evenly distributed in time and frequency, and therefore the TFR based features average out the noise quite effectively. The TD features,

which are based on time-series statistics, are quite vulnerable to the effects of noise.

Although there is no reason to suspect that additive noise is a reasonable model of the intra-class variance of transient MES datasets, these results corroborate that it is not. Class separability performs as well or better than PCA, and the STFT features are clearly the most effective. This is in contrast to the superiority of PCA-based feature sets and the WPT when classifying the transient MES.

5.3.2 Temporal Translation

The acquisition of transient MES patterns requires some form of automatic triggering mechanism to frame the waveform. Applying an arbitrary trigger (such as an amplitude threshold, which has been used here) to a signal that has a highly irregular waveform suggests some discrepancy as to where the waveform really ought to begin. This may have the effect of introducing temporal dispersion into a set of triggered patterns. It is likely therefore that temporal translation contributes, in part, to the intra-class variance in the transient MES datasets.

To investigate this effect in isolation, four template patterns were used to create an artificial dataset as in the additive noise example. The template from each class was subject to a random temporal shift, and replicates were generated to form the training, validation and test sets. The shift was uniformly distributed over a range $[-L, L]$, where L is defined to be a maximum shift limit. As L was varied from 0 to

20 samples⁴, negligible error was introduced into the test set classification error using the PCA-reduced WT. It is explained in Appendix D that the shift-induced modification of the WT coefficients manifests itself as additive random noise. Moreover, it is demonstrated in this appendix that PCA tends to relegate the shift-induced “noise” to the lesser PCA coefficients. Therefore, a shift of the template patterns not does introduce a significant amount of intra-class variance into the leading principal components.

To determine the interaction of shift with a real dataset, the two channel data were subject to artificially imposed random shift. For each subject, the random shift was uniformly distributed over $[-L, L]$, and L was varied from 0 to 20. The test set classification error, averaged across all subjects, is shown in Figure 5.14.

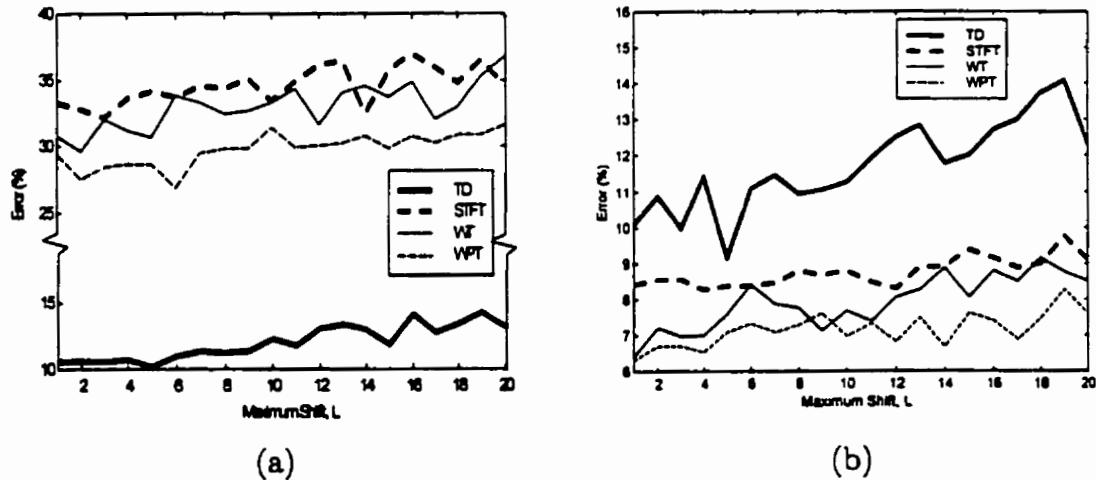


Figure 5.14 – The effect of temporal shift upon the test set error of the real transient MES, averaged across all subjects. Figure (a) shows the results when using CS, and (b) when using PCA dimensionality reduction. A LDA classifier has been used here; the results when using a MLP are essentially the same.

When acting upon real data, a random shift does degrade the test set classification error with increasing effect as L increases. All feature sets are affected by shift,

⁴It is unreasonable to expect that the jitter due to triggering would exceed this range.

but the TFR-based sets are less affected when using PCA as opposed to CS (experiencing a total degradation of 1-2% instead of 4-6%). This is due to the ability of PCA to suppress the shift-induced “noise” in the leading PCA features. Due to the relative insensitivity of the performance to shift when using PCA, it may be concluded that this is not a major source of intra-class variance. The lack of sensitivity of the PCA-reduced WPT to shift adds credence to the speculation that the WPT/PCA signal representation may be yielding results that are close to the Bayes error.

Curiously, the performance of the CS-reduced WT and WPT features degrade when subject to shift, but to a degree no worse than the CS-reduced TD and STFT feature sets. Moreover, the degradation is on the order of 4-6% error, whereas the absolute error for the unshifted data is roughly 30%. This suggests that accommodating the variance due to the loose structure of the MES places greater demands on the signal representation than accommodating any dispersion due to shift.

5.3.3 Superposition of Motor Groups

It has been shown that the surface recorded MES may contain the superimposed activity of many motor units, and possibly, multiple muscles [Hudgins91][Zip78]. The number of motor units and the degree of “cross-talk” between muscles depends upon the geometry of electrode placement and the volume conduction properties of the interstitial tissue.

To construct a simplified model of the surface MES, it will be assumed that there are “groups” of motor units that exhibit a similar discharge pattern. This concept has been referred to as the “common-drive” theory [DeLuca96]. To further simplify this model, assume that the discharge pattern associated with each “motor group” is consistent amongst contractions. In this situation, the motor group will elicit a deterministic “macro” pattern that is determined by the neural discharge pattern and the shape of the constituent motor units. Each of these motor groups will affect the recording in a proportion determined by the degree to which they are recruited. The variance in this model is based on the assumption that the recruitment scheme may vary for a given contraction type.

Consider a four-class problem in which the MES activity is assumed to comprise the activity of four motor groups. If we denote the activity of each motor group as s_1, s_2, s_3 , and s_4 , then we may model the measured signal as

$$x = As_1 + Bs_2 + Cs_3 + Ds_4 \quad (5.3)$$

where A, B, C and D are random variables. If we are to provide some distinction amongst the classes, the relative magnitudes A, B, C and D must be class dependent. The following model was used to assign the relative contribution of each motor group in a normally distributed manner:

$$\begin{aligned} \text{Class 1: } & A = N(1,.33) \quad B = N(.75,.25) \quad C = N(.75,.25) \quad D = N(.5,.15) \\ \text{Class 2: } & A = N(.75,.25) \quad B = N(1,.33) \quad C = N(.5,.15) \quad D = N(.75,.25) \\ \text{Class 3: } & A = N(.5,.15) \quad B = N(.75,.25) \quad C = N(.75,.25) \quad D = N(1,.33) \\ \text{Class 4: } & A = N(.75,.25) \quad B = N(.5,.15) \quad C = N(1,.33) \quad D = N(.75,.25) \end{aligned} \quad (5.4)$$

where $N(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ . The means and standard deviations were chosen to provide some probabilistic

overlap amongst classes. The patterns s_1, s_2, s_3 , and s_4 were chosen to represent the activity of individual motor groups. These are the localized activity at the sites of the short and long head of the biceps brachii (s_1 and s_2), and the short and long head of the triceps brachii (s_3 and s_4), measured using closely spaced bipolar electrodes⁵. These patterns are shown in Figure 5.15.

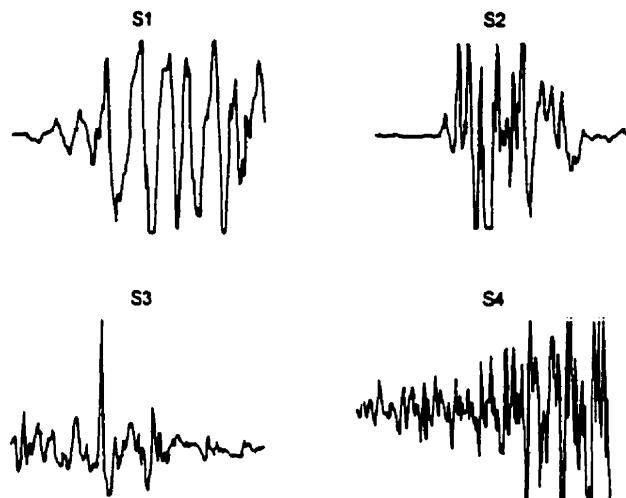


Figure 5.15 – The measured patterns used to represent the activity of individual motor groups. The signals roughly correspond to the activity of the short and long heads of the biceps brachii (s_1 and s_2), and the short and long heads of the triceps brachii (s_3 and s_4).

Using these signals as templates, training, validation and test sets were generated according to the random superposition model given above. The test set classification error was determined using each feature set, dimensionality reduction technique, and classifier. The results are shown in Figure 5.16.

⁵ These measurements are not sufficiently localized to be deemed motor groups; this would require an intra-muscular recording. They do, however, constitute patterns of different motor unit composition and firing rates, which is sufficient for this model.

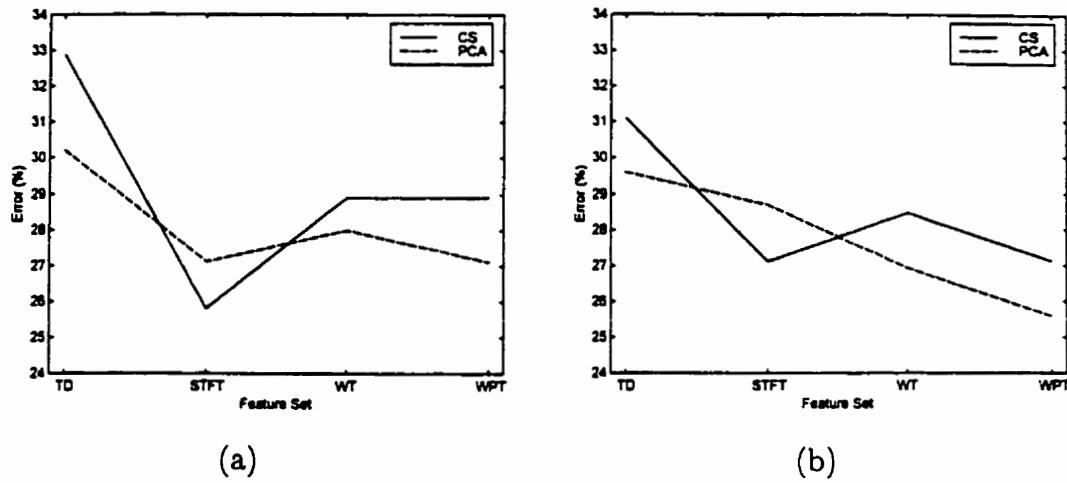


Figure 5.16 – The test set classification error of the superposition-based dataset. Figure (a) shows the results of CS and PCA based dimensionality reduction when using a LDA classifier, and (b) when using a MLP classifier.

The TFR based features sets perform better than the TD feature set, demonstrating that this problem requires discrimination in the time-frequency domain. The important characteristic however, is that there is very little difference between the results obtained when using CS and PCA. If this model were to accurately model the intra-class variance, PCA should show substantially better performance. Presumably, its deficiency is that it does not incorporate variance due to motor unit discharge patterns.

Aside: Triangular Waveform Dataset with Known Bayes Error.

It is insightful to compare the MES superposition problem to a similar problem with a known Bayes error. A waveform recognition problem was described by Breiman *et al.* [Breiman84]. It is a three-class problem based on the triangular waveforms $h_1[i], h_2[i], h_3[i]$ of length $n = 32$ such that

$$\begin{aligned} h_1[i] &= \max(6 - |i - 7|, 0) \\ h_2[i] &= h_1[i - 8] \\ h_3[i] &= h_1[i - 4] \end{aligned} \tag{5.5}$$

These waveforms are graphed in Figure 5.17.

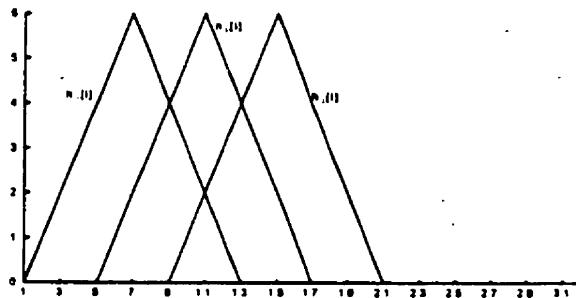


Figure 5.17 – Triangular waveforms.

Three classes of signals were generated as a random convex combination of two of these waveforms with additive noise:

$$\begin{aligned}
 \text{Class 1 : } & x^{(1)}[i] = uh_1[i] + (1-u)h_2[i] + \epsilon[i] \\
 \text{Class 2 : } & x^{(2)}[i] = uh_1[i] + (1-u)h_3[i] + \epsilon[i] \\
 \text{Class 3 : } & x^{(3)}[i] = uh_2[i] + (1-u)h_3[i] + \epsilon[i]
 \end{aligned} \tag{5.6}$$

where u is a uniform random variable on the interval $(0,1)$ and the $\epsilon[i]$ are standard normal deviates. It has been shown [Breiman84] that the Bayes error for this problem is about 14%.

A training set of 252 patterns and a test set of 750 patterns contained an equal number of signals from each class. Each of the TFR feature sets was applied to this problem. The TD feature set was not used, since the time statistics are meaningless on a record of such short duration. The STFT used a window length of 8 and an overlap of 50% to accommodate the short record length. Figure 5.18 depicts the test set classification error as a function of feature set dimension.

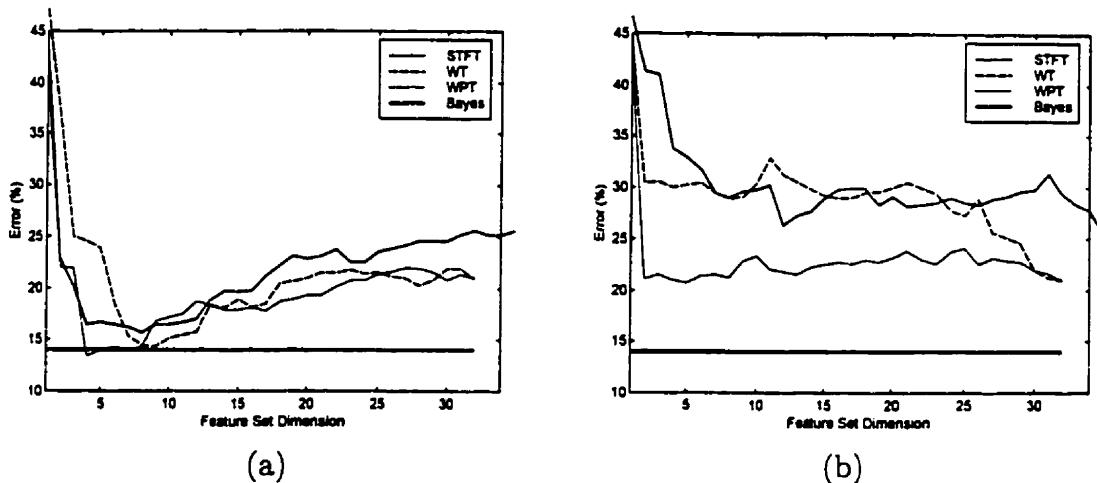


Figure 5.18 – The test set classification error of the triangular waveform data as a function of feature set dimension. Figure (a) shows the results when using CS feature selection, and (b) when using PCA feature projection.

The most striking effect is that CS outperforms PCA. This is the opposite of what has been observed throughout this work when analyzing the transient MES. The CS-reduced features perform very well, with the best performance when using between 4-10 features. There is a gradual improvement from STFT to WT to WPT; the WPT approaches (and matches, at a dimension of 4) the Bayes error, which is the best that one could expect, on average. This model, like that which generates a random contribution of motor groups, fails to explain the intra-class variance of the transient MES.

The models of intra-class variance comprising *additive noise*, *temporal translation* and *superposition of motor groups* impose some effect that acts upon the entire waveform. It is possible to construct other “macro structure” models of intra-class variance, such as amplitude/frequency modulation, or the injection of other sources of noise. None of these methods however, can explain the structural variance of the MES and consequently, the superiority of PCA to CS dimensionality reduction. An appropriate model of this phenomenon must be based upon the processes that constitute the local structure of the waveforms. The next section presents such a model.

5.3.4 Wavelet Transform Decomposition Model

At the onset of sudden muscular effort, there appear to be neuromuscular schemata that produce patterns of motor unit recruitment and neural discharge timing. This activity has sufficient order to yield surface MES waveforms that are visually similar amongst repeated trials of a given effort. A slight modification of the recruitment or discharge scheme however, can greatly modify the local characteristics of the MES waveform. A human observer can perceive similarities amongst an ensemble of waveforms corresponding to a particular class due to the ability of the visual system to associate local patterns as belonging to a higher level perceptual structure⁶. Conveying this loosely defined structure is a challenging task for statistical signal representation methods.

The evidence presented in this work indicates that the intra-class variance in the transient MES is predominantly due to the loose temporal structure of the waveforms. To model this intra-class variance then, we must consider the neuromuscular processes from which the signals originate. It is certainly possible to construct a model of the transient MES. An ensemble of motor unit action potentials (MUAPs) of varying shape, a recruitment scheme, and neural discharge statistics can be defined, and some random behavior could be introduced to model the intra-class variance. It would be very difficult to have these patterns resemble real MES waveforms however, since very little is known about recruitment schemes and firing patterns during rapid contractions. Motor unit decomposition methods do exist, but these are successful only with records of constant or slowly changing force production [Stashuk88]. With typical motor unit firing rates (10-20 Hz), there is only time for 2-5 discharges in the interval of the transient MES

⁶ This is the ultimate goal of syntactic signal representation and pattern recognition.

patterns considered here. With the suggestion that an orderly scheme of recruitment and firing is taking place, it is possible that a mean firing rate is meaningless in the context of short, ballistic bursts of activity.

5.3.4.1 Modeling the Intra-Class Variance

An empirical approach to modeling the transient MES is proposed here. One of the primary strengths of the wavelet transform is its ability to localize activity in time and in frequency. This is embodied in the wavelet basis functions; each translated and scaled version of the mother wavelet seeks a correlation with the signal at a specific location and scale. Of particular interest, many wavelet functions “look like” typical MUAPs. What this suggests is that a wavelet decomposition provides an estimate of the elemental behavior of the transient MES at different locations and scales. It does this without assuming a model of recruitment or firing rate. Although it does not explicitly identify the activity of individual motor units, it does provide an effective characterization of the local structure in the waveform⁷.

Consider a typical MES burst pattern, as shown in Figure 5.19. This is the activity of the biceps during elbow flexion. The signal was subject to a wavelet decomposition using a Coiflet-4 wavelet, primarily because it resembles the shape of a typical triphasic MUAP. Each wavelet coefficient in the decomposition corresponds to the activity at a particular location and scale. To demonstrate this, the top four coefficients (in terms of amplitude) were retained, and were

⁷ As well, the high-scale (low-frequency) coefficients can incorporate the effects of motion artifact or mechanical motion of the constituent muscle groups.

individually subject to the WT reconstruction algorithm. The portion of the signal that each basis function is modeling is shown in Figure 5.19, superimposed on the original waveform.

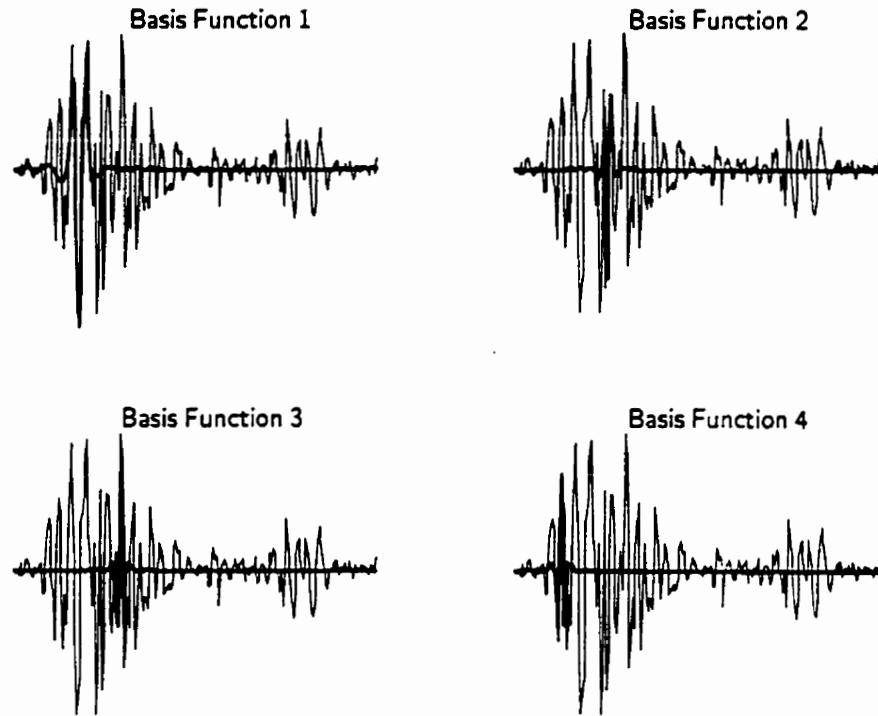


Figure 5.19 – The reconstruction of the four largest WT coefficients. The location and the scale of the reconstructed signals are determined by the basis function; the amplitude of the response is specified by the magnitude of the corresponding WT coefficient.

With the notion that the WT coefficients are modeling the local structure within a pattern, it is possible to accommodate modification to this local structure. Specifically, the main effect of variations in MUAP recruitment and firing rates is to alter the magnitude and timing of the constituent motor units. The wavelet model can emulate this by amplitude modulation and translation of the WT coefficients.

To model an arbitrary pattern, a template of real MES is used to perform a full WT decomposition, as above. Variation in the timing of the local structure is

achieved by allowing each WT coefficient to shift by a random amount. Similarly, assigning a random scale factor to each WT coefficient accommodated the variation in amplitude of the components of local structure. Using each coefficient $w_{j,k}$ in the WT decomposition, a modified set of coefficients is produced:

$$\tilde{w}_{j,k+\tau} = \alpha w_{j,k} \quad (5.7)$$

where j is the scale, k is the location, and α, τ are uniform random variables describing amplitude and shift variations. Notice that the new set of coefficients $\{\tilde{w}_{j,k}\}$ may contain a superposition of multiple coefficients because the re-indexing $k \rightarrow k + \tau$ is a random map. The ranges of α and τ are specified by assigning maximum limits to the uniform distributions. The range of α , limited to a maximum scale of A , is defined to produce a uniform distribution on the interval $[\frac{1}{A}, A]$. The range of τ , limited to a maximum shift of T , is defined to produce a uniform distribution on the interval $[-T, T]$.

Figure 5.20 demonstrates the ability of this model to generate the variations in temporal structure present in the transient MES. The five patterns on the left represent real MES patterns corresponding to biceps activity during elbow flexion. The five patterns on the right have been generated using the WT model, using a maximum shift of $T = 5$ and a maximum amplitude scale of $A = 4$.

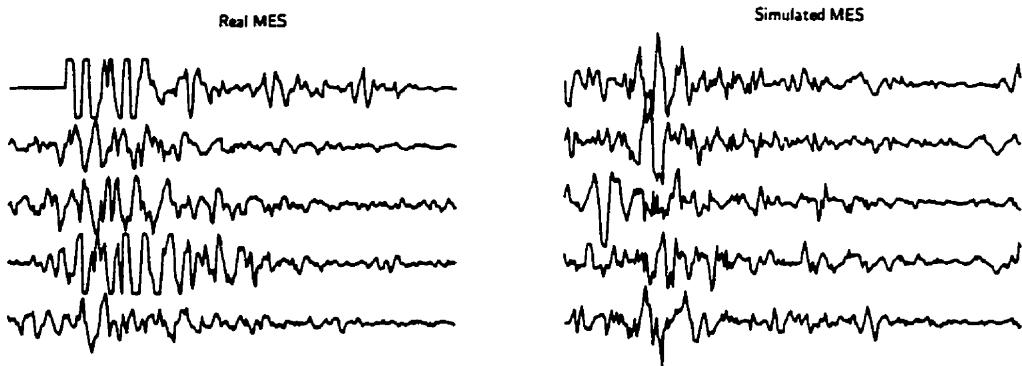


Figure 5.20 – Ensembles of real (left) and simulated (right) transient MES patterns. The simulated patterns are derived from the top pattern of real MES, using a maximum coefficient shift of five, and a maximum amplitude scale of four.

It is visually apparent that the model is doing a credible job of simulating the structural variations present in the sequence of real MES patterns.

It must now be determined if datasets generated using this WT model place the same demands on a classification system as does the real MES. From an arbitrary subject, template patterns were taken from each of four classes of contraction for both biceps and triceps channels. From these templates, replicates were generated to yield a dataset equal in size to the datasets of real data (100, 150, 150 patterns distributed amongst the training, test and validation sets respectively). In each case, the validation set was used to estimate the optimal feature set dimension.

To use this model most effectively, the effect of the magnitude of shift, T , and amplitude scale, A , must be determined. First, the maximum amplitude scale was fixed at $A = 2$, and the magnitude of shift was varied from $T = 1$ to $T = 20$. The effect upon test set classification error when using CS and PCA dimensionality reduction is shown in Figure 5.21.

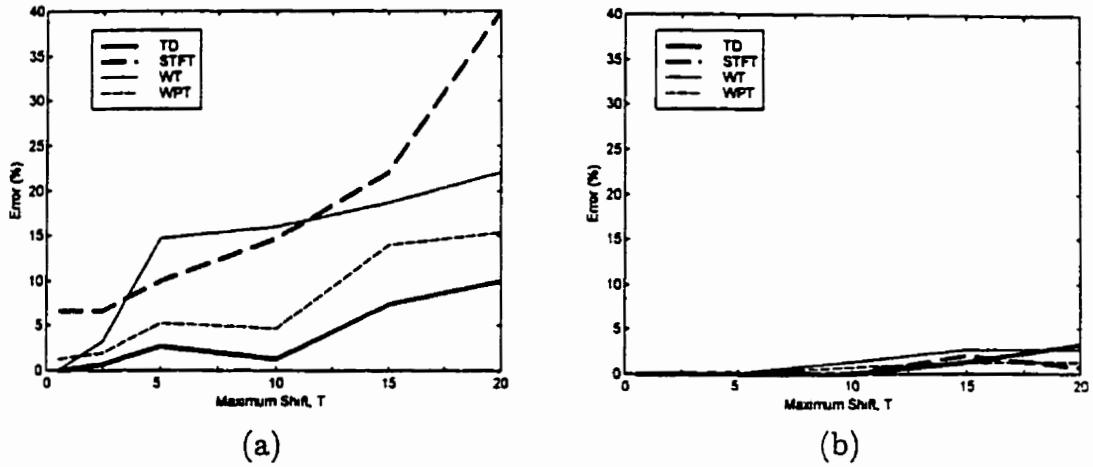


Figure 5.21 – The effect of the maximum range of shift in the WT model. The test set classification error is shown for each feature set when using (a) CS, and (b) PCA dimensionality reduction. These results were obtained using a LDA classifier; the results when using a MLP were essentially the same.

Next, the range of shift was fixed at $T = 10$, and the limits of amplitude scale were varied from $A = 1$ to $A = 6$. The effect upon test set error is shown in Figure 5.22.

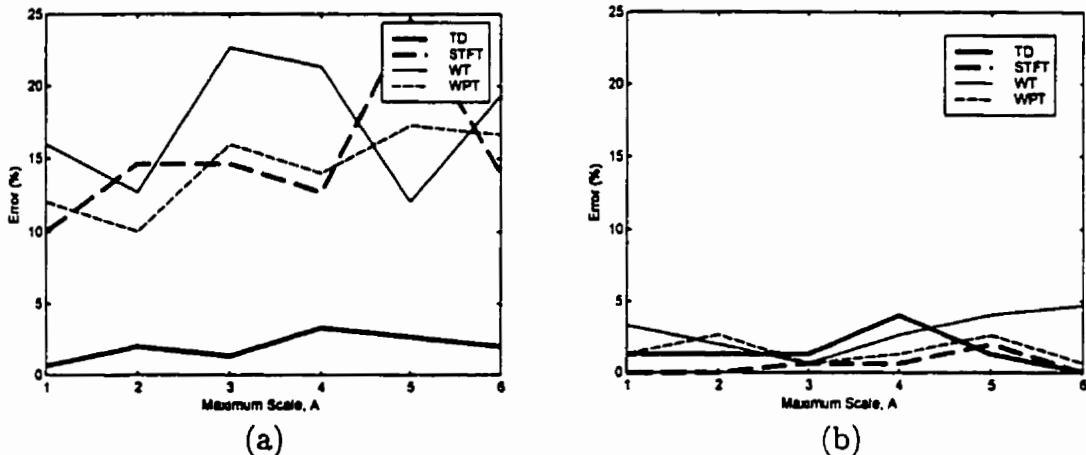


Figure 5.22 – The effect of the maximum range of amplitude scale in the WT model. The test set classification error is shown for each feature set when using (a) CS, and (b) PCA dimensionality reduction. These results were obtained using a LDA classifier; the results when using a MLP were essentially the same.

In Figure 5.21 and Figure 5.22, PCA clearly outperforms CS dimensionality reduction, as is the situation with real MES data. The classification performance degrades with increasing range of both shift and amplitude scale, although shift has a much more pronounced effect. The goal of this parameter selection is to produce a dataset that exhibits the same degree of complexity (with respect to

classification performance) as the real MES data. Although somewhat arbitrary, a WT model with a maximum shift of $T = 15$ and a maximum amplitude scale of $A = 4$ seems to be a justifiable choice.

Having established an appropriate set of parameters for the WT model, the performance of each feature set and dimensionality reduction technique was determined. Figure 5.23 shows the test set classification error upon a dataset generated by the WT model.

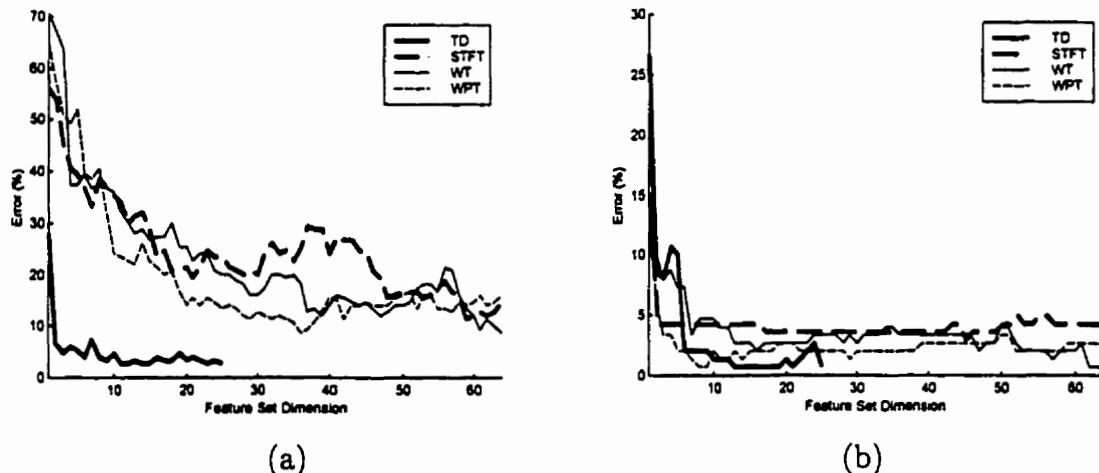


Figure 5.23 – The effect of feature set dimension upon the test set classification error of the simulated transient MES. The results are shown for (a) CS, and (b) PCA dimensionality reduction. The results have been generated using a LDA classifier; a MLP yields essentially the same characteristics.

The response is similar in many ways to that which would be expected of a dataset of real MES. The performance of PCA is clearly superior to that of CS when using TFR based features. This is presumably due to the dispersion of information in the time-frequency domain, which is apparent upon inspection of the TFRs of the simulated dataset. The relationship between classification error and feature set dimension is typical of that exhibited by a real dataset, for both CS and PCA. The relative performance of the datasets is difficult to determine from a single dataset, especially when using PCA features because the classification error is so

small. Using this model, it is not possible to generate a dataset that experiences PCA error that is substantial. This implies that a significant portion of PCA error in real datasets is due to effects that are not of neuromuscular origin, such as system error or operator error. If this model does provide an accurate representation of the intra-class variance, it provides further evidence that the performance is near the Bayes bound.

5.3.4.2 Bootstrapping Using Simulated MES Data

The utility of the WT model goes beyond its ability to explain the structural variations and consequently, the requirements of the signal representation. Because the data are derived from pattern templates of a real dataset, it is conceivable that the simulated patterns could serve to bootstrap this dataset. This is especially important in clinical situations where it is difficult to acquire a training set of sufficient size to fully train a classifier.

To demonstrate this, the training sets from each subject in the two channel database were decimated from 100 patterns to 33. It has been observed that, for the data from most subjects, between 50-80 patterns are needed to adequately train the LDA and MLP classifiers. Using these depleted training sets, the test set classification rate was determined. Next, the training data were bootstrapped using the WT model. For each real training set pattern, two simulated patterns were generated and added to the training set, such that the bootstrapped training set consisted of 33 real and 66 simulated patterns. The test set classification error

(averaged across all subjects) when using the original (full) training set, the depleted training set, and the bootstrapped training set is shown in Figure 5.24.

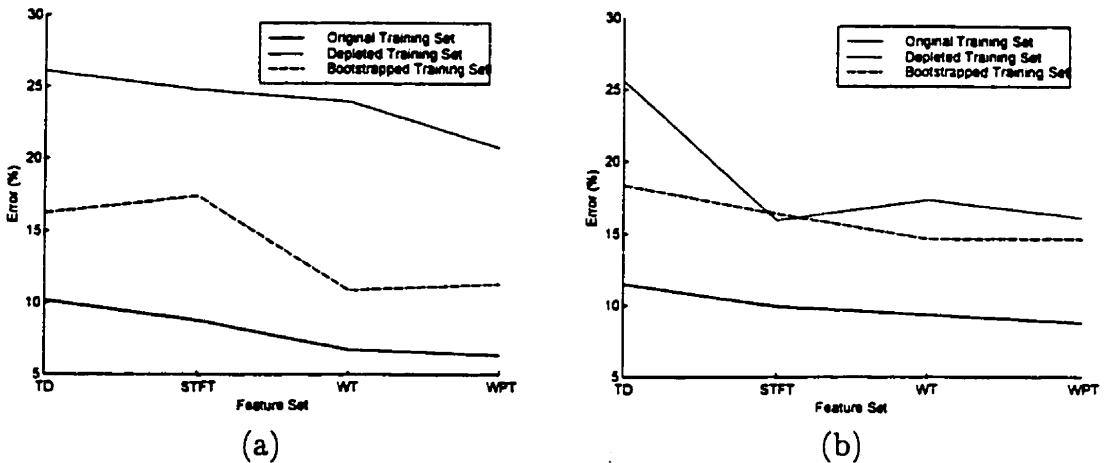


Figure 5.24 – The effect of bootstrapping the training set using the WT model. The test set classification error, averaged across all subjects, is shown when using the original training set (100 patterns), a depleted training set (33 patterns), and a bootstrapped training set (the depleted set plus 66 simulated patterns). The response is shown for PCA-reduced TD, STFT, WT, and WPT features when using (a) a LDA, and (b) a MLP classifier.

For both classifiers and for each feature set, the bootstrapping training set improves the test set generalization. This indicates that the simulated patterns do convey some information about the intra-class variance that exists in the real data.

The need for a larger training set often arises when acquiring data from a limb-deficient individual. The remaining musculature may be sparse or of low tone, and fatigue may prevent the acquisition of sufficient data. This was the situation with an individual with an above-elbow amputation, labeled subject ‘O’ in Hudgins’ work [Hudgins91]. Only 80 patterns were collected, and 40 were assigned to each of the training and test sets. Under the assumption that this training set is insufficient to fully train a classifier, the data were bootstrapped using the WT model. For every real training set pattern, a simulated pattern was generated, creating a training set size of 80. The test set classification error for this subject

when using the original training set and the bootstrapped set is shown in Figure 5.25.

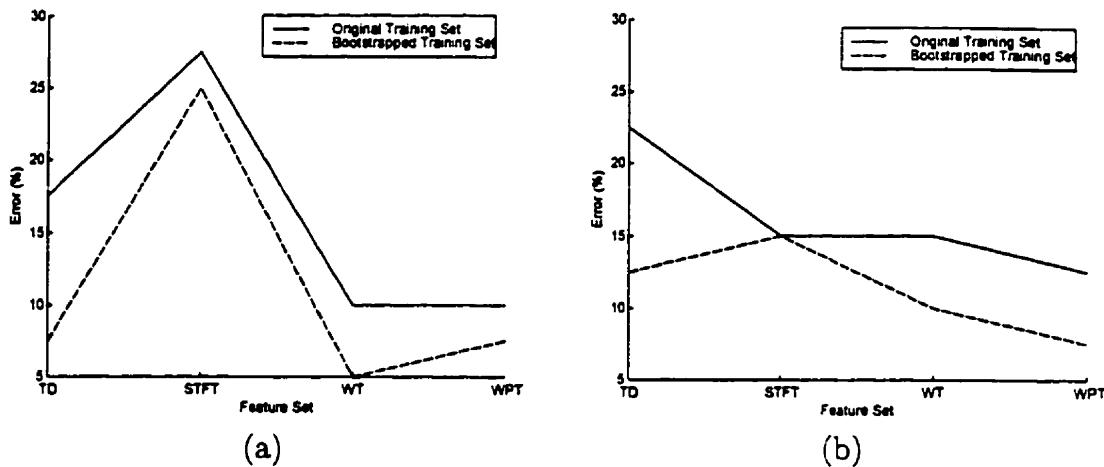


Figure 5.25 – The effect of bootstrapping a sparse training set, acquired from a limb-deficient individual. The test set classification error is shown for each feature set when using (a) a LDA, and (b) a MLP classifier.

Bootstrapping has the desired effect, allowing an improvement over that possible when using the original training set. As well, it is apparent that WT and WPT features outperform the TD feature set. It should be noted that the variability in these results may be high, due to the relatively small test set.

5.4 Summary

This chapter has provided a closer look at the classification performance of the transient MES. In doing so it has described 1) methods of facilitating the clinical implementation, 2) the bounds on performance, and 3) a model of why the problem is a challenging one.

It has been established that a generalized dimension may be used in lieu of determining the dimension on an individual basis, without sacrificing performance. This simplifies the process of acquiring data and determining the dimensionality of the signal representation. PCA-reduced TFR feature sets are relatively insensitive to dimensionality specification within a dimensionality range of 20-50. When using the WPT/PCA combination, a generalized dimension of 30 will match the performance of an individually specified dimension.

It has been suggested that the best results observed may be near the Bayes error. Several methods have been proposed which should reduce the error beyond that of the WPT/PCA representation, but in fact do not. It seems very likely that a significant portion of the observable misclassification is due to system error and operator error.

Several models of intra-class variance were investigated. Variability due to additive noise, temporal translation, and the superposition of gross patterns of muscle activity fail to explain the demands placed upon the scheme of signal representation. The implication is that the intra-class variance is not due to macro effects, but rather, the elemental processes which constitute the waveform. A

model based upon the WT was used to characterize the local structure of the transient MES. Datasets generated by this model appear to reproduce the structural variation inherent in real MES patterns, and exhibit the same dependency upon dimensionality reduction methods and feature set dimension. The simulated data has been shown to effectively augment sparse training sets of real data.

Chapter 6

Conclusions

6.1 Summary

The purpose of this thesis was to investigate the role of signal representation in the classification of transient myoelectric signal patterns. The objective was to provide a better understanding of the factors that influence the efficacy of signal representation for classification and consequently, to provide improved classification accuracy.

Chapter 1 introduced the etiology of the patterns of transient MES activity. Whereas the temporal waveform of the surface MES has been considered to be devoid of structure, this work exploits Hudgins' observation that loose structure does exist in the patterns that accompany the onset of sudden contractions. The neuromuscular processes from which these patterns may originate are described, and some rationale for the structure is developed. An examination of these patterns suggests that this structure is manifest in the time-frequency domain, and that the information is distributed in a complex manner. These observations suggest that 1) a time-frequency representation would provide an effective

characterization of these patterns, and 2) some means of extracting the relevant information from the complex time-frequency characteristics is required.

In Chapter 2, signal representation for pattern classification was described in detail. The problem was broken down into the tasks of feature extraction, dimensionality reduction, and classification. The stages of feature extraction and dimensionality reduction comprise the procedure of signal representation. It is emphasized that, although an appropriate classifier is necessary, it is the signal representation that profoundly affects the classification performance of a given problem. In a real-time application, it is essential that the feature extraction have an efficient implementation. In this regard, one is limited to linear TFRs as candidate time-frequency transforms. Correspondingly, the STFT, WT and WPT representations were introduced. The importance of dimensionality reduction in classification was described, and alternative methods were characterized as belonging to one of two methodologies: *feature selection* or *feature projection*. The relative performance of these two approaches provides important insight throughout the thesis.

Chapter 3 provided the conceptual and mathematical background of wavelet theory. The wavelet transform was introduced first, and an extension was made to accommodate wavelet packet transforms. Methods of specifying an orthonormal basis from the redundant WPT were described, explaining that the original methods of basis selection were motivated by signal compression. A recent modification to WPT basis selection – the local discriminant basis (LDB) algorithm – was described as a means of deriving feature sets for classification. The LDB is fundamental to the use of the WPT in this work.

Chapter 4 presented the classification results of TFR feature sets and dimensionality reduction strategies. The focus of these analyses centered upon a database of two channel transient MES patterns, acquired from 16 normally-limbed subjects. For each feature set (TD, STFT, WT, and WPT) the optimal transform parameters and dimensionality reduction strategies were determined empirically, based upon the ensemble of two channel datasets. For all TFR feature sets, feature projection (in the form of PCA) was shown to dramatically outperform feature selection based dimensionality reduction techniques. PCA was also shown to be superior to commonly used *ad hoc* methods of dimensionality reduction. Moreover, an improvement due to the relative efficacy of feature sets has been demonstrated in the progression TD → STFT → WT → WPT. Overall, the best performance was obtained when using the WPT/PCA signal representation, yielding an average classification error of 6.25%.

Chapter 5 provided some perspective on the results of Chapter 4. It was demonstrated that a PCA based feature set is relatively insensitive to feature dimension. This allows a generalized dimension to be specified for all subjects, obviating the time-consuming process of collecting sufficient data to constitute a validation dataset.

An ideal assessment of the efficacy of the WPT/PCA representation would be in comparison to a Bayes classifier, which is an elusive task. The Bayes error is a nebulous quantity, which assumes the existence of a probabilistic model of the data, or an infinite ensemble of exemplars drawn from the same distribution. The data here are neither infinite nor identically distributed, so the Bayes error cannot be estimated. A general observation has been made however, that the average classification performance of the two channel data may be approaching the Bayes

bound. To do so, several methods of improving the performance of the signal representation were proposed. While these methods did improve the signal representations that performed poorly, none could improve upon the performance of the WPT/PCA combination. After factoring in operator error and system error, the average classification error of the WPT/PCA representation leaves very little room for improved discrimination of the “good” patterns. Chapter 5 also included an investigation of the intra-class variance that characterizes the transient MES classification problem. Global effects such as additive noise, temporal translation and superposition of macro patterns were shown to be inadequate models. A model based upon a WT decomposition of the signals was proposed with the intention of estimating the neuromuscular processes that give rise to the local structure of the waveforms. The injection of random behavior into this model was shown to provide an effective emulation of the intra-class variance of real datasets.

Above all, the objectives of this thesis have been met. The wavelet packet transform has been shown to outperform all other feature sets under consideration, averaging 6.25% error. This represents a significant improvement over Hudgins' method, which, for these data, yields an average error of 9.25%. It has been shown that, when using TFR based feature sets, PCA provides a far more effective means of dimensionality reduction than feature selection by CS. Moreover, by preprocessing the feature set with PCA prior to classification, a LDA – a classifier that is easier to implement and to train than a MLP – may be used without degrading performance.

The improved accuracy provided by these methods will enhance the functionality of a pattern recognition based myoelectric control system. The WPT has a

complexity on the order of $n \log n$, and the stages of PCA and LDA, once trained, require a simple matrix multiplication in the feedforward path. Therefore, the combination WPT/PCA/LDA easily lends itself to real-time implementation on a DSP microprocessor of modest capabilities.

6.2 Original Contributions

In the author's opinion, the original contributions of this work are:

1. The methods investigated in this work offer improved classification performance of the transient MES in a computationally efficient manner. The average error of 6.25% when using the WPT/PCA combination represents a significant improvement over Hudgins' TD feature set, which, for these data, yields an average error of 9.25%.
2. It has been demonstrated that feature projection (PCA) based dimensionality reduction outperforms a variety of feature selection methods when using TFR feature sets to classify the transient MES. Moreover, the classification performance is relatively insensitive to feature set dimension, obviating the need to determine the optimum dimension *via* a validation set.
3. Performing PCA dimensionality reduction simplifies the task of the classifier to the extent that a LDA can be used in the place of a MLP without degrading performance. The advantages of using a LDA are i) it is conceptually simple and easily interpreted, ii) it trains very quickly for the datasets used here, iii) it does not require any fine-tuning of its architecture or training algorithm. These factors make the LDA an attractive choice in a clinically-oriented application.
4. The parameters of the TD, STFT, WT and the WPT feature sets have been empirically optimized to give the best possible classification performance for

the two channel MES. A gradual improvement in performance has been demonstrated in the progression TD→STFT→WT→WPT when using PCA dimensionality reduction. To the author's knowledge, this is the first successful application of wavelet and wavelet packet based features to transient MES classification.

5. It has been conjectured that the average performance of the WPT/PCA combination is close to the Bayes performance for the ensemble of data acquired for this work. Empirical evidence was gathered and presented to support this conjecture.
6. A model of the transient MES has been proposed based upon a WT decomposition of the signal. This model has been shown to emulate the structural variation in the transient MES and consequently, the relative performance of feature sets and dimensionality reduction techniques. This WT model has also been shown to provide an effective means of bootstrapping datasets that are of insufficient size to fully train a classifier. This is especially valuable when acquiring data from limb-deficient individuals, who are often incapable of enduring extended data collection sessions.

6.3 Future Work

Improvements in Accuracy

It is possible that other methods of feature extraction and dimensionality reduction may prove to be superior to the methods investigated here. New methods of signal representation are being proposed at an unprecedented rate.

1. Shift-invariant WT and WPT transforms, as described in Section 4.4, may offer improved performance. Although true shift-invariance has yet to be offered (while preserving orthogonality and computational efficiency), this is a topic of enormous interest. Other time-frequency representations exist that may prove advantageous in classification. The matching pursuit method [Zhang93] provides a redundant basis that seeks to explain the signal variance. It has been shown to be immune to the effects of shift experienced by the WT and the WPT [Blinowska94]. Quadratic TFRs promise high-resolution representations, but are limited by substantial computational complexity and problems with cross-term interference. Progress has been made on both counts [Qian96], but quadratic TFRs are not yet feasible in real-time applications.
2. Dimensionality reduction by feature projection has been performed using PCA in this work, but other methods exist that offer more complex objective functions such as projection pursuit [Huber85] and independent component analysis [Karhunen97]. Other projection methods allow nonlinear projections, such as nonlinear PCA [Kramer91] and Kohonen maps [Kohonen89]. It is possible that these methods may outperform PCA in classification applications.

3. Committees of classifiers have been subject to recent interest in the literature. Combining multiple estimators, each of which exploit different types and sets of features, may yield better performance than the best single estimator. Crucial to the success of these methods is how to properly combine them; that is, how to integrate the information from the ensemble of classifiers [Jordan94]. Other classification paradigms should be considered. Genetic algorithms have been shown to be capable of learning very complex problems. The evolutionary learning procedure amounts to a form of feature extraction by a process of elimination [Koza92]. Preliminary work in MES classification has shown promising results [Farry97].

Improvements in Prosthetic Function

4. The degree of success with a four-class problem suggests extending the problem to more classes. How does classification performance degrade as one tries to discriminate five, six, or more classes? A reliable seven-state system (six movement classes plus one inactive state) would allow some states to be defined as combined motions designed to fulfill active daily living tasks (such as an elbow/wrist linkage, for feeding motions).
5. A greater understanding of the incidence and influence of training errors is necessary. An experiment should be designed in an attempt to identify operator error and system error (as defined in Section 5.2). What proportion of the overall error rate is due to operator error and system error? What is the error rate when these patterns are removed? How severely is the classifier's performance being compromised by having to "learn" mistakes in the training set? Would a more sophisticated triggering mechanism reduce the incidence of operator and system error?

6. The present approach requires generating MES patterns from a state of inactivity. This is a significant constraint in myoelectric control as many activities become cumbersome if a period of rest is required when switching devices. A substantial improvement would be realized with the ability to identify patterns of intent in a background of extraneous activity. This requires a much more sophisticated method of framing the patterns than the level threshold scheme used in this work. The classification system must act as a detector as well as a classifier, implying that the data must be processed in a temporally continuous stream. This imposes a challenging problem at the stages of signal representation and classification. A dynamic neural network has been applied to the task of continuous-mode detection/classification of transient underwater sounds, with some success [Ward98].
7. If a many-class continuous classification system is feasible, it may be advantageous to abandon the concept of discrete classes or "states", in favor of a set of primitives that define a more fundamental unit of MES activity. This now becomes a syntactic model, requiring a grammar which establishes interrelationships amongst the primitives. This is the foundation of continuous speech recognition, which has become a commercial entity in 1998. The most successful speech recognition systems use Hidden Markov Models which learn the detailed relationships amongst the primitives of speech. The analogue of continuous speech recognition in prosthetic control would be continuous and independent control of each device. It is not clear however, that the dynamic behavior of the MES is sufficiently structured to support such a syntactic model.

References

[Akay95]

Akay, M., "Wavelets in biomedical engineering," *Annals of Biomedical Engineering*, vol. 23, pp. 531-542, 1995.

[Albus79]

Albus, J.S., "Mechanisms of planning and problem solving in the brain," *Mathematical Biosciences*, 45, pp. 247-293, 1979.

[Almstrom81]

Almstrom, C., Herberts, P., and L. Korner, "Experience with Swedish multifunction prosthetic hands controlled by pattern recognition of multiple myoelectric signals," *Int. Orthopaed.* 5, 15-21, 1981.

[Atsma97]

Atsma, W.J., *Classification of Myoelectric Signals using Neural Networks*, M.Sc. Thesis, University of New Brunswick, Fredericton, N.B., Canada, 1997.

[Auscher92]

Auscher, P., Weiss, G., and M.V. Wickerhauser, *Local Sine and Cosine Bases of Coifman and Meyer and the Construction of Smooth Wavelets*, in: *Wavelets: a Tutorial in Theory and Applications* (C.K. Chui, ed.), Academic Press, pp. 237-256, 1992.

[Baldi89]

Baldi, P. and K. Hornik, "Neural networks and principal component analysis: learning from examples without local minima," *Neural Networks*, 2, (1), pp. 53-58, 1989.

[Bartnik92]

Bartnik, E.A. and K.J. Blinowska, "Wavelets - new method of evoked potential analysis," *Medical and Biological Engineering and Computing*, 30, pp. 125-126, 1992.

[Basmajian85].

Basmajian, J. and C.J. DeLuca, *Muscles Alive*, Baltimore: Williams and Wilkins, 5th Ed., 1985.

[Basseville89]

Basseville, M., "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, no. 4, pp. 349-369, 1989.

[Beck94]

Beck, S. and L. Deuser, "Automatic classification of acoustic sequences by multiresolution image processing and neural networks," *IEEE Int. Conference on Image Processing*, Vol. 3, pp. 931-935, 1994.

[Bellman61]

Bellman, R., *Adapted Control Processes: A Guided Tour*, Princeton University Press, 1961.

- [Bentley98]
Bentley, P.M., Grant, P.M., and J.T.E. McDonnell, "Time-frequency and time-scale techniques for the classification of bioprosthetic valve sounds," *IEEE Transactions on Biomedical Engineering*, vol. 45, No. 1, January, 1998.
- [Beylkin92]
Beylkin, G., "On the representation of operators in bases of compactly supported wavelets," *SIAM J. Numerical Analysis*, vol. 29, No. 6, pp. 1716-1740, 1992.
- [Bienenstock82]
Bienenstock, E., Cooper, L., and P. Munro, "Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex." *Journal of Neuroscience*, vol. 2, pp. 32-48, 1982.
- [Bigland[54]]
Bigland, B. and O.C.J. Lippold, "The relation between force, velocity and integrated electrical activity in human muscles", *J. Physiol.*, vol. 123, pp. 214-224, 1954.
- [Bishop95]
Bishop, C.M., "Training with noise is equivalent to Tikhonov regularization," *Neural Computation*, 7(1): 108-116, 1995.
- [Bishop96]
Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1996.
- [Blinowska94]
Blinowska, K.J., and P.J. Durka, "The application of wavelet transform and matching pursuit of the time-varying EEG signals," *Intelligent Engineering Systems Through Artificial Neural Networks*. Vol. 4, Ed. Dagli, Fernandez, Gosh, pp. 535-540, 1994.
- [Boashash89]
Boashash, B. and P.J. Black, "An efficient real-time implementation of the Wigner-Ville Distribution," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 35, No. 11., pp. 1611-1818, November, 1987.
- [Boashash90]
Boashash, B. and P. O'Shea, "A methodology for detection and classification of some underwater acoustic signals using time-frequency analysis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 38, No. 11, pp. 1829-1841, November, 1990.
- [Bourlard88]
Bourlard, H. and Y. Kamp, "Auto-association by multilayer perceptrons and singular value decomposition," *Biological Cybernetics*, 59, pp. 291-294, 1988.
- [Bovik92]
A.C. Bovik, N. Gopal, T. Emmoth and A. Restrepo, "Localized measurement of emergent image frequencies by Gabor wavelets," Special Issue on Wavelet Transforms and Multiresolution Signal Analysis, *IEEE Transactions on Information Theory*, vol. IT-38, no. 3, pp. 691-712, March 1992.

- [Brieman84]
Breiman, L., Friedman, H., Olsen, R.A. and C.J. Stone, *Classification and Regression Trees*, Chapman and Hall, Inc., New York, 1984.
- [Bruce96]
Bruce, A., Donoho, D., and H.-Y. Gao, "Wavelet analysis," *IEEE Spectrum*, pp. 26-35, October, 1996.
- [Buchtal54]
Buchtal, F., Pinelli, P. and P. Rosenflack, "Action potential parameters in normal human muscle and their physiological determinants", *Acta. Physiol. Scand*, vol. 32, pp. 219-229, 1954.
- [Carpenter87]
Carpenter, G.A. and S. Grossberg, "A massively parallel architecture for a self-organizing neural pattern recognition machine," *Computer Vision, Graphics, and Image Processing*, 37: pp. 54-115, 1987.
- [Casasent94a]
Casasent, D.P. and J.-S. Smokelin, "Neural net design of macro Gabor wavelet filters for distortion-invariant object detection in clutter," *Optical Engineering*, 33(7). pp.2264-2271, 1994.
- [Casasent94b]
Casasent, D.P. and J.-S. Smokelin, "Real, imaginary, and clutter Gabor filter fusion for detection with reduced false alarms," *Optical Engineering*, 33(7), pp.2255-2263, 1994.
- [Childress69]
Childress, D.A., "A myoelectric three state controller using rate sensitivity," in *Proc. 8th ICMBE*, Chicago, IL, 1969, S4-5.
- [Chui94]
Chui, C.K. Shi, X. and A.K. Chan, "An oversampled frame algorithm for real-time implementation and applications," *Proc. Of SPIE Wavelet Applications Conference*, vol. 2242, pp. 272-301, 1994.
- [Choi94]
Choi, H.G., Principe, J.C., Hutchinson, A.A. and J.A. Wozniak, "Multiresolution segmentation of respiratory electromyographic signals," *IEEE Trans. Biomedical Engineering*, vol. 41, No. 3, pp. 257-266, 1994.
- [Cohen95]
Cohen, I., Raz, S., and D. Malah, "Shift invariant wavelet packet bases," *Proc. of ICASSP*, vol. 2, pp. 1081-1084, 1995.
- [Cohen89]
Cohen, L., "Time-frequency distributions - A Review," *Proc. IEEE*, vol. 77, No. 7, pp. 941-981, July, 1989.

- [Cohen90]
Cohen, A., Daubechies, I. And J.C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Communications of Pure and Applied Math*, 1990.
- [Coifman89]
Coifman, R.R. and Y. Meyer, *Nouvelles bases orthonormées de $L^2(\mathbb{R})$ ayant la structure du système de Walsh*, Dept. Mathematics, Yale University, New Haven, CT. Aug., 1989.
- [Coifman90]
Coifman, R.R. and Y. Meyer, "Orthonormal wave packet bases", Dept. Mathematics. Yale University, New Haven, CT. 1990.
- [Coifman91]
Coifman, R.R. and Y. Meyer, "Remarques sur l'analyse de Fourier à fenêtre." *Comptes Rendus, Acad. Sci., Paris. Serie. I*, 312, pp. 259-261, 1991.
- [Coifman92]
Coifman, R.R. and M.V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE. Trans. Information Theory*, vol. 38, No. 2, pp. 713-719, 1992.
- [Coifman94]
Coifman, R.R. and M.V. Wickerhauser, "Adapted waveform analysis as a tool for modeling, feature extraction, and denoising," *Optical Engineering*, 33(7), pp. 2170-2174. 1994.
- [Coifman95]
Coifman, R.R. and D.L. Donoho, "Translation-invariant de-noising," in: *Wavelets and Statistics*, (Eds. A. Antoniadis and G. Oppenheim), Lecture Notes in Statistics. Springer-Verlag, pp. 125-150, 1995.
- [Cooley65]
Cooley, J.W. and J.W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Math. Comp.*, Vol. 19, pp. 297-301, 1965.
- [Daubechies88]
Daubechies, I., "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, 41, pp. 909-996, 1988.
- [Daubechies92]
Daubechies, I., *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 61, SIAM, Philadelphia, 1992.
- [Daubechies96]
Daubechies, I., "Where do wavelets come from - a personal point of view," *Proc. IEEE*. vol. 84, N.o 4, pp. 510-513, 1996.
- [DelMarco94]
DelMarco, S. and J. Weiss, "M-band wavepacket-based transient signal detector using a translation-invariant wavelet," *Optical Engineering*, vol. 33, No. 7, pp. 2175-2182. 1994.

[DelMarco96]

DelMarco, S. and J. Weiss, "Improved transient signal detection using a wavepacket-based detector with an extended translation-invariant wavelet transform," *Optical Engineering*, vol. 35, No. 1, pp. 131-137, 1996.

[DeLuca96]

DeLuca, C.J., Foley, P.J., and Z. Erim, "Motor unit control properties in voluntary isometric isotonic contractions," *Journal of Neurophysiology*, 76, 1503-1516, 1996.

[DeLuca79]

DeLuca, C.J.. "Physiology and mathematics of myoelectric signals," *IEEE Trans. Biomed. Eng.*, 26, 313-325, 1979.

[Denker91]

Denker, J.S. and Y. le Cun, "Transforming neural-net output levels to probability density functions." In R.P. Lippmann, J.E. Moody, and D. Touretsky, editors, *Advances in Neural Processing Systems 3*, pp. 853-859, Morgan Kaufmann, 1991.

[Devlin81]

Devlin, S.J., Gnanadesikan, R. and J.R. Kettenring, "Robust estimation of dispersion matrices and principal components," *J. American Statistical Association*, v. 76, pp. 354-362, 1981.

[Dickhaus96]

Dickhaus, H. and H. Heinrich, "Classifying biosignals with wavelet networks," *IEEE Engineering in Medicine and Biology*, pp. 103-113, September/October, 1996.

[Dickhaus94]

Dickhaus, H. Khadra, L. and J. Brachmann. "Time-frequency analysis of ventricular late potentials, *Methods Inform. Med.*, 33(2), pp. 197-195, 1994.

[Doershuk83]

Doerschuck, P.C., Guftafson, D.E., and A. S. Willsky, "Upper extremity limb function discrimination using EMG signal analysis," *IEEE Trans. Biomed. Eng.*, 30, 18-28, 1983.

[Dorcas66]

Dorcas, D. and R.N. Scott, "A three state myoelectric control," *Medical and Biological Engineering*, 4, 367-372, 1966.

[Drosopoulos92]

Drosopoulos, A. and S. Haykin, "Adaptive radar estimation with Thompson's multiple window method," *Adaptive Radar Estimation*, Haykin and Steinhardt, editors. John Wiley and Sons, Inc, 1992.

- [Duda73]
Duda, R.O. and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, NY, 1973.
- [Durand90]
Durand, L.-G., Blanchard, M., Cloutier, G., Sabbah, H.N., and P.D. Stein, "Comparison of pattern recognition methods for computer-assisted classification of heart sounds in patients with a porcine bioprosthetic valve implanted in the mitral position," *IEEE Trans. Biomed. Eng.*, vol. 37, pp. 1121-1129, 1990.
- [Durka96]
Durka, P., *Time-Frequency Analyses of EEG*, PhD. Dissertation, Warsaw University. Warsaw, Poland, August, 1996.
- [Elman90]
Elman, J., "Finding structure in time," *Cognitive Science*, vol. 14, pp. 179-211, 1990.
- [Etemad94]
Etemad, K. and R. Chellappa, "Separability based tree structured local basis selection for texture classification," *Proc. IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, 1994.
- [Farry97]
Farry, K., Fernandez, J., Abramczyk, R., Novy, M., and D. Atkins, "Applying genetic programming to control of an artificial arm," *Myoelectric Controls Conference 1997*, Fredericton, N.B., Canada, pp. 50-55, 1997.
- [Farry96]
Farry, K., Walker, I.D., and R.G. Baraniuk, "Myoelectric teleoperation of a complex robotic hand," *IEEE Trans. Robotics and Automation*, 12 (5), pp. 775-788, 1996.
- [Fisher36]
Fisher, R.A., "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, pp. 179-188, Reprinted in *Contributions to Mathematical Statistics*, John Wiley, New York, 1950.
- [Fisher36]
Fisher, R.A., "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, Vol. 7, pp. 179-188, 1936.
- [Fourier88]
Fourier, J.B.J., "Théorie analytique de la chaleur," in *Oeuvres de Fourier*, tome premier, G. Darboux, Ed., Paris: Gauthiers-Villars, 1888.
- [Frazer94]
Frazer, G. and B. Boashash, "Multiple window spectrogram and time-frequency distributions," *Signal Processing Research Centre*, QUT, P.O. Box 2434, Brisbane, 4001, Australia.
- [Friedman74]
Friedman, W.J. and J.W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Transactions on Computers*, Vol. 23, pp. 881-889, 1974.

- [Fu82]
Fu, K.S., *Syntactic Pattern Recognition and Applications*, Prentice-Hall, Englewood Cliffs, N.J., 1982.
- [Fukunaga90]
Fukunaga, K., *Introduction to Statistical Pattern Recognition 2nd Ed.*, Academic Press. San Diego, CA, 1990
- [Gabor46]
Gabor, D., "Theory of communication," *J. Inst. Elec. Eng.*, 93, 429-457, 1946.
- [Gallant93]
Gallant, P.J., "An Approach to Myoelectric Control Using a Self-Organizing Neural Network for Feature Extraction," Master's Thesis, Queen's University, Kingston, Ontario. 1993.
- [Georgopoulos93]
Georgopoulos, A. et al. "Cognitive neurophysiology of the motor cortex," *Science*. Vol. 260, pp. 47-52, 1993.
- [Ghosh90]
Ghosh J., Deuser, L. and S. Beck, "Impact of feature vector selection on static classification of acoustic transient signals," *Gov. Neural Network Applications Workshop*, Aug., 1990.
- [Ghosh92]
Ghosh, J., Deuser, L., and S. Beck, "A neural network based hybrid system for detection, characterization, and classification of short-duration oceanic signals," *IEEE Journal of Oceanic Engineering*, vol. 17, No. 4, October, pp. 351-363, 1992.
- [Giles94]
Giles, C.L., Kuhn, G.M. and R.J. Williams, "Dynamic recurrent neural networks: theory and applications," *IEEE Transactions on Neural Networks*, vol. 5 No. 2, pp. 153-155, 1994.
- [Gopinath91]
Gopinath, R.A. and C.S. Burrus, *Wavelets and Filter Banks*, in: *Wavelets: A Tutorial in Theory and Applications*, ed. C.K. Chui, Academic Press, 1991.
- [Graupe82]
Graupe, D., Salahi, J., and K.H. Kohn, "Multifunction prosthesis and orthosis control via microcomputer identification of temporal pattern differences in single-site myoelectric signals," *J. Biomed. Eng.*, 4, 17-22, 1982.
- [Grossmann84]
Grossmann, A. and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM Journal of Mathematical Analysis*, 15(4), pp. 723-736, July, 1984.

[Guglielmi96]
Guglielmi, R., *Wavelet Feature Definition and Extraction for Classification and Image Processing*, Ph.D Thesis, Yale University, New Haven CT, 1996.

[Gyaw94]
Gyaw, T.A. and S.R. Ray, "The wavelet transform as a tool for recognition of biosignals," *Biomedical Sciences and Instrumentation*, vol. 30, pp. 63-68, 1994.

[Hanson89]
Hanson, S.J. and L.Y. Pratt, "Comparing biases for minimal network construction with backpropagation," In D. Touretsky, editor, *Advances in Neural Information Processing Systems 1*, pp.177-185, Morgan Kaufmann, 1989.

[Hart68]
Hart, P.E., "The condensed nearest neighbor rule," *IEEE Transactions on Information Theory*, 114, pp. 515-516, 1968.

[Haykin94]
Haykin, S., *Neural Networks: A Comprehensive Foundation*, Maxwell MacMillan Canada, Inc., Don Mills, Ontario, 1994.

[Hecht-Neilson90]
Hecht-Nielson, R., *Neurocomputing*, Reading, MA: Addison-Wesley, 1990.

[Hemminger94]
Hemminger, T.L. and P. Yoh-Han, "Detection and classification of underwater acoustic transients using neural networks, *IEEE Trans. Neural Networks*, vol. 5, No. 5. pp. 712-718, 1994.

[Henneman65]
Henneman, E., Somjen, G. and D.O. Carpenter, "Functional significance of cell size in spinal motoneurons", *J. Physiol.*, vol. 28, pp. 560-580, 1965.

[Hess-Neilsen95]
Hess-Neilsen, N, and M.V. Wickerhauser, "Wavelets and time-frequency analysis," *Proc. IEEE*, vol. 84, No. 4, pp. 523-540, 1996.

[Hicks93]
Hicks, C.R., *Fundamental Concepts in the Design of Experiments*. Saunders College Publishing, New York, 1993.

[Hlawatsch92a]
Hlawatsch, F. and P. Flandrin, "The interference structure of the Wigner distribution and related time-frequency distributions," in *The Wigner Distribution – Theory and Applications in Signal Processing*, W. Mecklenbraucker, ed., North Holland Elsevier Science Publishers, 1992.

- [Hlawatsch92b]
Hlawatsch, F. and G.F. Boudreux-Bartels, "Linear and Quadratic Time-Frequency Representations," *IEEE Signal Processing Magazine*, pp. 21-67, 1992.
- [Holmstrom97]
Holmstrom, L., Koistinen, P., Laaksonen, J., and E. Oja. "Neural and statistical classifiers – Taxonomy and two case studies," *IEEE Trans. Neural Networks*, Vol. 8, No. 1, January, 1997.
- [Hopfield82]
Hopfield, J.J., "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Science USA*. 79. pp. 2554-2558, 1982.
- [Huber85]
Huber, P.J., "Projection pursuit (with discussion)," *The Annals of Statistics*, vol. 13. No. 2, pp. 435-475, 1985.
- [Hudgins93]
Hudgins, B., Parker, P.A., and R.N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans. Biomedical Engineering*, 40, 1, 82-94, 1993.
- [Hudgins91]
Hudgins, B.S., *A New Approach to Multifunction Myoelectric Control*. Ph.D. Dissertation, University of New Brunswick, Fredericton, N.B., Canada, 1991.
- [Hush93]
Hush, D.R. and B.G. Horne, "Progress in supervised neural networks: What's new since Lippmann," *IEEE Signal Processing Magazine*, pp. 8-39, January, 1993.
- [Huynh98]
Huynh, Q., Cooper, L., Intrator, N. and H. Shouval, "Classification of underwater mammals using feature extraction based on time-frequency analysis and BCM theory," *To appear: IEEE Transactions On Signal Processing, Special issue on NN*.
- [Intrator97a]
Intrator, N., Huynh, Q., and G. Dobeck, "Feature extraction from acoustic backscattered signals using wavelet dictionaries," *Proceedings of SPIE97*, April, 1997.
- [Intrator97b]
Intrator, N., Wong, Y., Huynh, Q. and B. Lee, "Wavelet feature extraction for discrimination tasks," *Proceeding of the 1997 Canadian Workshop in Information Theory*, Toronto, June, 1997.
- [Intrator92]
Intrator, N. and L.N. Cooper, "Objective function formulation of the BCM theory of visual cortical plasticity: statistical connections, stability considerations," *Neural Networks*, vol. 5, pp. 3-17, 1992.

[Intrator91]

Intrator, N. and G. Tajchman, "Supervised and unsupervised feature extraction from a cochlear model for speech recognition," in *Neural Networks for Signal Processing - Proceedings of the 1991 Workshop* (B. Juang, S. Kung and C. Kamm, eds.), pp. 460-469. IEEE Press, 1991.

[Jacobs88]

Jacobs, R.A., "Increased rates of convergence through learning rate adaptation," *Neural Networks*, Vol. 1, No. 4, pp. 295-308, 1988.

[Jones91]

Jones, D.L. and R.G. Baraniuk, "Efficient approximation of the continuous wavelet transform," *Electronics Letters*, 27, 748-750, 1991.

[Jordan94]

Jordan, M.I. and R.A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Computation*, vol. 3, pp. 181-214, 1994.

[Jordan89]

Jordan, M., "Serial order: a parallel, distributed processing approach," *Advances in Connectionist Theory: Speech*. Hillsdale: Lawrence Erlbaum, 1989.

[Kadambe94]

Kadambe, S. and P. Srinivasan, "Applications of adaptive wavelets for speech," *Optical Engineering*, 33(7), pp. 2204-2211, 1994.

[Kalayci95]

Kalayci, T. and O. Ozdamar, "Wavelet preprocessing for automated neural network detection of EEG spikes," *IEEE Engineering in Medicine and Biology*, pp. 160-165. March/April, 1995.

[Karhunen47]

Karhunen, K. "Über linearen methoden in der wahrscheinlichkeitsrechnung," *Ann. Acad. Sci. Fennicae*, Ser. A, 37, no. 1, 1947.

[Karhunen97]

Karhunen, J., Oja, E., Wang, L., Vigario, R., and J. Joutsensalo, "A class of neural networks for independent component analysis," *IEEE Trans. Neural Networks*, Vol. 8, No. 3, May 1997.

[Karhunen95]

Karhunen, J. and J. Joutsensalo, "Generalizations of principle component analysis. optimization problems, and neural networks," *Neural Networks*, vol. 8, No. 4, pp. 549-562. 1995.

[Kelly90]

Kelly, M. and P.A. Parker, "The application of neural networks to myoelectric signal analysis: A preliminary study," *IEEE Trans. Biomedical Engineering*, BME-37, No. 3. pp. 221-230, 1990.

[Kohonen88]

Kohonen, T., "An introduction to Neural Computing," *Neural Networks*, 1(1), pp. 3-16. 1988.

[Kohonen89]

Kohonen, T., *Self Organization and Associative Memory*, Berlin: Springer-Verlag, 3rd ed., 1989

[Kovacevic92]

Kovacevic, J. and M. Vetterli, "Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for R^n ," *IEEE Trans. Information Theory*. vol. 38, No. 2, 1992

[Koza92]

Koza, J.R., *Genetic Programming: On the Programming of Computer by Natural Selection*, Cambridge, MA: MIT Press, 1992.

[Kramer91]

Kramer, M.A., "Nonlinear principal component analysis using autoassociative neural networks," *AIChe Journal*, 37 (2), pp. 233-243, 1991.

[Kruskal69]

Kruskal, J.B., "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation'," *Statistical Computation*, (R.C Milton and J.A. Nelder, eds.). Academic Press, New York, pp. 427-440, 1969.

[Kuehner97]

Kuehner, N., Madsen, P., and H. Kunov., "Adaptive wavelet packet based stimulus artifact removal in transiently evoked otoacoustic emission testing", *Canadian Medical and Biological Engineering Society Conference*, pp. 30-31, Toronto ON, May, 1997.

[Kullback51]

Kullback, S. and R.A. Leibler, "On information and sufficiency," *Annals of Mathematics and Statistics*, 22, pp. 79-86, 1951.

[Kundu94]

Kundu, A., Chen, G.C., and C.E. Persons, "Transient sonar signal classification using hidden Markov models and Neural Nets," *IEEE Journal of Oceanic Engineering*. vol. 19, No. 1, pp. 87-99, 1994.

[Kuruganti95]

Kuruganti, U., Hudgins, B. and R.N. Scott, "Two-channel enhancement of a multifunction control scheme," *IEEE Trans. Biomedical Engineering*, Vol. 42, No. 1. January 1995.

[le Cun90]

le Cun, Y., Denker, D.S. and S.A. Solla, "Optimal brain Damage," In D. Touretsky, editor, *Advances in Neural Information Processing Systems 2*, pp. 598-605. Morgan Kaufmann, 1990.

[Learned95]

Learned, R.E. and A.S. Willsky, "A wavelet packet approach to transient signal classification," *Applied and Computational Harmonic Analysis*, vol. 2, No. 3, pp. 256-278, 1995.

[Lee94]

Lee, S.-Y. and H. Szu, Fractional Fourier transforms, wavelet transforms, and adaptive neural networks, *Optical Engineering*, 33(7), pp.2326-2330, 1994.

[Li93]

Li, C., Zheng, C. and T. Changfeng, "Detection of ECG characteristic points using wavelet transforms," *IEEE Trans. Biomedical Engineering*, vol. 42, No. 1, 1992.

[Liang96]

Liang, J., and T.W. Parks, "A translation-invariant wavelet representation algorithm with application," *IEEE Trans. Signal Processing*, vol. 44, No. 2, pp. 225-232, 1996.

[Lim95]

Lim, L.M., Akay, M. and A.J. Daubenspeck, "Identifying respiratory-related evoked potentials," *IEEE Engineering in Medicine and Biology*, pp. 174-178, March/April, 1995.

[Lin96]

Lin, Z and J. Chen, "Advances in time-frequency analysis of biomedical signals," *Critical Reviews in Biomedical Engineering*, 24(1), pp. 1-76, 1996.

[Lindstrom77]

Lindstrom, L. and R. Magnusson, "Interpretation of myoelectric power spectra: a model and its applications", *Proc. IEEE*, vol. 65, pp. 653-662, 1977.

[Lippmann87]

Lippmann, R., "An introduction to computing with neural nets," *IEEE Acoustics, Speech and Signal Processing Magazine*, Vol. 4, No. 2, pp. 4-22, April, 1987.

[Liu97]

Liu, B. and S.-F. Ling, "Tree-structured local basis selection using genetic algorithms," Preprint in *Wavelet Digest* (<http://www.wavelet.org/cm/ms/what/wavelet/>), December, 1996.

[LoConte94]

Lo Conte, L.R., Merletti, R. and G.V. Sandri, "Hermite expansions of compact support waveforms: applications to myoelectric signals," *IEEE Trans. Biomedical Engineering*, Vol. 41, pp. 1146-1159, 1994.

[Lowe91]

Lowe, D. and A. Webb, "Optimized feature extraction and the Bayes decision in feed-forward classifier networks," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 13, pp. 355-364, Apr., 1991.

[Majid93]

Majid, F., Coifman, R.F. and M.V. Wickerhauser, "The XWPL system reference manual," Yale University, 1993.

- [Makhoul91]
Makhoul, J., "Pattern recognition properties of neural networks," *Proc. 1st IEEE Workshop on Neural Networks for Signal Processing*, pp. 173-187, Sept. 1991.
- [Mallat89a]
Mallat, S., "Multiresolution approximations and wavelet orthonormal bases in $L^2(\mathbb{R})$," *Trans. Amer. Math. Soc.*, 315, pp. 69-87, 1989.
- [Mallat89b]
Mallat, S., "A theory for multiresolution signal decomposition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 11, No. 2, pp. 617-643, 1988.
- [Mallat91]
Mallat, S., "Zero-crossings of a wavelet transform," *IEEE Trans. on Information Theory*, vol. 37, No. 4, pp. 1019-1033, 1991.
- [Mallat92a]
Mallat, S. and S. Zhong, "Characterization of signals from multiscale edges," *IEEE Trans. Pattern Recognition and Machine Intelligence*, vol. 14, No. 7, 1992.
- [Mallat92b]
Mallat, S. and S. Zhong, "Wavelet transform maxima and multiscale edges," in: *Wavelets and Their Applications*, ed: M.B. Ruskai, Jones and Bartlett, Boston, 1992.
- [Mallat93]
Mallat, S. and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, Vol. 41, pp. 3397-3415, Dec. 1993.
- [Malvar90]
Malvar, H.S., "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoustics, Speech and Signal Processing*, 38, pp.969-987, 1990
- [Marinovic85]
Marinovic, N.M. and G. Eichmann, "Feature extraction and pattern classification in space-spatial frequency domain," *SPIE vol. 579 Intelligent Robots and Computer Vision*. pp. 19-25, 1985.
- [Marple87]
Marple, S.L., *Digital Spectral Analysis with Applications*, Prentice-Hall, Englewood Cliffs, New Jersey, 1987.
- [Matsuoka94]
Matsuoka, M. and M. Kawamoto, "A neural network that self-organizes to perform three operations related to principle component analysis," *Neural Networks*, vol. 7, No. 5. pp. 753-765, 1994.
- [Meyer93]
Meyer, Y., *Wavelets and Operators*, Cambridge Studies in Advanced Mathematics, vol. 37, Cambridge Univ. Press, New York, Translated by D.H. Salinger, 1993.

- [Meyer93]
Meyer, Y., *Wavelets: Algorithms and Applications*, SIAM, Philadelphia, PA., Translated and Revised by R.D. Ryan, 1993.
- [Miller90]
Miller, A.J., *Subset Selection in Regression*, Chapman & Hall, 1990.
- [Milner73]
Milner-Brown, H.S., Stein, R.B. and R. Yemm, "Changes in firing rate of human motor units during linearly changing voluntary contractions", *J. Physiol.*, 230, p. 371, 1973.
- [Moody89]
Moody, J. and C.J. Darken, "Fast learning in networks of locally-tuned processing units," *Neural Computation*, 1, pp. 281-293, 1989.
- [Nason94]
Nason, G.P. and B.W. Silverman, "The discrete wavelet transform in S," *Journal of Computational and Graphical Statistics*, vol. 3, pp. 163-191, 1994.
- [Nason95]
Nason, G.P. and B.W. Silverman, "The stationary wavelet transform and statistical applications," in: *Wavelets and Statistics*, (Eds. A. Antoniadis and G. Oppenheim). Lecture Notes in Statistics, Springer-Verlag, pp. 281-299, 1995.
- [Oja89]
Oja, E., "Neural networks, principal components, and subspaces," *International Journal of Neural Systems*, 1 (1), pp. 61-68, 1989.
- [Paliwal91]
Paliwal, K.K., "A time-derivative neural network architecture - an alternative to the time-delay neural net architecture." In *Neural Networks for Signal Processing - Proceedings of the 1991 IEEE Workshop*, eds. B.H. Juang, S.Y. Kung, and Candace A. Kamm, pp. 367-375, 1991.
- [Papoulis65]
Papoulis, A. *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill, New York, 1965.
- [Parker86]
Parker, P.A., and R.N. Scott, "Myoelectric control of prostheses," *CRC Critical Reviews in Biomedical Engineering*, 13, 4, 283-310, 1986.
- [Person72]
Person, R.S. and L.P. Kudina, "Discharge frequency and discharge frequency in human motor units during voluntary contractions of muscle", *Electromyog. clin. Neurophysiol.*, pp. 471-483, 1972.
- [Pesquet96]
Pesquet, J.C., Krim, H. and H. Carfantan, "Time invariant orthonormal wavelet representations," *IEEE Trans. Signal Processing*, vol. 44, No. 8, pp. 1964-1970, 1996.

[Piche94]

Piche, S.W., "Steepest descent algorithms for neural network controllers and filters," *IEEE Transactions on Neural Networks*, vol. 5 No. 2, pp. 198-212, 1994.

[Pineda88]

Pineda, F.J., "Dynamics and architecture for neural computation," *Journal of Complexity*, 4, pp. 216-245, 1988.

[Qian96]

Qian, S. and D. Chen, *Joint Time-Frequency Analysis: Methods and Applications*, Prentice-Hall, Toronto, 1996.

[Rosenblatt58]

Rosenblatt, F., "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychological Review*, 65, pp. 386-408, 1958.

[Rioul93]

Rioul, O., "Regular wavelets: a discrete-time approach," *IEEE Trans. Signal Processing*, Vol. 41, No. 12, pp. 3572-3579, 1993.

[Saito93]

Saito, N. and G. Beylkin, "Multiresolution representations using thre auto-correlation functions of compactly supported wavelets," *IEEE Trans. Signal Processing*, Vol. 41, pp. 3584-3590, Dec. 1993.

[Saito94a]

Saito, N., *Local Feature Extraction and its Applications using a Library of Bases*. Ph.D. Thesis, Dept. of Mathematics, Yale University, New Haven, CT USA, December. 1994.

[Saito94b]

Saito, N., *Wavelets in Geophysics*, E. Foufoula-Georgiou and P. Kumar (eds.), Academic Press, Inc., New York.

[Saito95]

Saito, N. and R.R. Coifman, "Local discriminant bases and their applications," *J. Math. Imaging and Vision*, vol. 5, No. 4, pp. 337-358, 1995.

[Saito96a]

Saito, N. and R. Coifman, "Improved local discriminant bases using empirical probability density estimation," *Amer. Statist. Assoc. Proc. Statistical Computing*, 1996.

[Saito96b]

Saito, N., "Classification of geophysical acoustic waveforms using time-frequency atoms," *Amer. Statist. Assoc. Proc. Statistical Computing*, 1996.

[Saito97]

Saito, N., Yale Univeristy / Doll-Schlumberger, *Personal Communication*, February, 1997.

[Sanger89]

Sanger, T.D. "Optimal unsupervised learning in a single layer feedforward neural network," *Neural Networks*, vol. 2, pp. 459-474, 1989.

- [Saridis82]
Saridis, G.N. and T.P. Gootee, "EMG pattern recognition for a prosthetic arm," *IEEE Trans. Biomed. Eng.*, **29**, 403-412, 1982.
- [Sarle97]
Sarle, W., USENET newsgroup moderator: comp.ai.neural-nets FAQ.
- [Sato90]
Sato, M., "A real time learning algorithm for recurrent analog neural networks," *Biological Cybernetics*, vol. 62, pp. 237-241, 1990
- [Schmeidl77]
Schmeidl, H., "The I.N.A.I.L. experience fitting upper-limb dysmelia patients with myoelectric control," *Bulletin of Prosthetics Research*, **10**, 27, 17-42, 1977.
- [Schmidt88]
Schmidt, R.A., *Motor Control and Learning, A Behavioral Emphasis*, Champaign: Human Kinetics, 1988.
- [Scott92]
Scott, D.W., *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley and Sons, New York, 1992.
- [Senhadji95]
Senhadji, L., Carrault, G., Ballanger, J.J. and G. Passariello, "Comparing wavelet transforms for recognizing cardiac patterns," *IEEE Engineering in Medicine and Biology*, pp. 167-172, March/April 1995.
- [Sietsma91]
Sietsma, J. and R.J.F. Dow, "Creating artificial neural networks that generalize," *Neural Networks*, **4**: 67-79, 1991.
- [Simoncelli92]
Simoncelli, E.P., Freeman, W.T., Adelson, E.H. and D.J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Information Theory*, vol. 38, No. 2, pp. 587-607, March, 1992.
- [Slepian61]
Slepian, D. and H.O. Pollak, "Prolate spheroidal wave functions, Fourier analysis and uncertainty – I," *Bell Systems Technical Journal*, Vol. 40, pp. 43-64, 1961.
- [Soteling94]
Soteling, L.G., Saerens, M. and H. Bersini, "Classification of temporal trajectories by continuous-time recurrent nets," *Neural Networks*, vol. 7, No. 5, pp. 767-776, 1994.
- [Specht90]
Specht, D.F., "Probabilistic neural networks," *Neural Networks*, **3**, pp. 109-118, 1990.

- [Stashuk88]
Stashuk, D. and H. DeBruin, "Automatic decomposition of selective needle-detected myoelectric signals," *IEEE Trans. Biomed. Eng.*, Vol. 35, pp. 1-10, 1988.
- [Steffen93]
Steffen, P., Heller, P., Gopinath, R.A. and C.S. Burrus, "Theory of regular M-band wavelet bases," *IEEE Trans. in Signal Processing*, December, 1993.
- [Stevenson93]
Stevenson, M., "Survey of dynamic neural network techniques with application to temporal processing tasks," *Defense Research Establishment Atlantic*, SSC file: OSC93-00782(011), 1993.
- [Strang97]
Strang, G., and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press. Wellesley, MA, 1997.
- [Sweldens96]
Sweldens, W., "Wavelets: What next?," *Proc. IEEE*, vol. 84, No. 4, pp. 680-688. 1996.
- [Szu92]
Szu, H.H., B. Telfer, and S. Kadambe, "Neural network adaptive wavelets for signal representation and classification." *Optical Engineering*, vol. 31, No. 9, pp. 1907-1916, 1992
- [Taswell93]
Taswell, C. and K.C. McGill, "Wavelet transform algorithms for finite-duration discrete-time signals," *Numerical Analysis Project Manuscript NA-91-07*, Department of Computer Science, Stanford University, 1993.
- [Taswell94]
Taswell, C., "Near-best basis selection algorithms with non-additive information cost functions," *IEEE-SP Symposium on TFTSA*, 1994.
- [Taswell95]
Taswell, C., "Top-down and bottom-up tree search algorithms for selecting bases in wavelet packet transforms," *Proceedings of the Villard de Lans Conference*, Springer Verlag, 1995.
- [Tate96]
Tate, A.R., *Pattern Recognition Analysis of In Vivo Magnetic Resonance Spectra*, Ph.D. Dissertation, University of Sussex at Brighton, September, 1996.
- [Telfer94]
Telfer, B.A., Szu, H.H., Garcia, J.P., Hanseok, K., Dubey, A. and N. Witherspoon, "Adaptive wavelet classification of acoustic backscatter and imagery," *Optical Engineering*, 33(7), pp. 2192-2203, 1994.
- [Thomson85]
Thomson, D., "Spectrum estimation and harmonic analysis," *Proc. IEEE*, Vol. 70, No. 9, pp. 1055-1096, Sept. 1982.

- [Tou74]
Tou, J.T. and R.C. Gonzalez, *Pattern Recognition Principles*, Addison-Wesley, Reading, MA, 1974.
- [Trejo93]
Trejo, L.J. and M.J. Shensa, "Linear and neural network models for predicting human signal detection performance from event-related potentials: A comparison of the wavelet transform with other feature extraction methods," *Proc. 5th Workshop on Neural Networks: SPIE Volume 2204, San Diego*, pp. 153-161, 1993.
- [Vidakovic]
Vidakovic, B. and P. Muller, *Wavelets for Kids*, Duke University Tech. Report.
- [Vodovnik67]
Vodovnik, L., Kreifeldt, J., Caldwell, R., Green, L., Silgalis, E., and P. Craig, "Some topics on myoelectric control of orthotic/prosthetic systems," *Rep. EDC 4-67-17*. Case Western Reserve University, Cleveland, OH, 1967.
- [Waibel89]
Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 37, No. 3, pp. 328-339, 1989.
- [Wantabe85]
Wantabe, S., *Pattern Recognition: Human and Mechanical*, John Wiley and Sons, New York, 1985.
- [Ward98]
Ward, M., *Detection and classification of transient underwater sounds by a finite impulse response neural network*, M.Sc. Thesis, University of New Brunswick. Fredericton, N.B., Canada, *in preparation*.
- [Warner96]
Warner, F., Yale University, *Personal Communication*, November, 1996.
- [Weiss93]
Weiss, J., "Translation-invariance and the wavelet transform," *Aware Technical Report* No. AD921102, 1993.
- [Wickerhauser94]
Wickerhauser, M.V., *Adapted Wavelet Analysis from Theory to Software*, AK Peters. Ltd., Wellesley, Massachusetts, 1994.
- [Wickerhauser94]
Wickerhauser, M.V., "Two fast approximate wavelet algorithms for image processing, classification, and recognition," *Optical Engineering*, 33(7), pp.2225-2235, 1994.
- [Widrow90]
Widrow, B. and M.A. Lehr, "30 years of adaptive neural networks: Perceptron, madaline, and backpropagation," *Proceedings of the IEEE*, 78, pp. 1415-1442, 1990.

[Wirta78]

Wirta, R.W., Taylor, D.R., and F.R. Findley, "Pattern recognition arm prosthesis: A historical perspective - Final Report," *Bulletin of Prosthetics Research*, **10**, 30, 8-35. 1978.

[Zhang95]

Zhang, J., G.G. Walter, Y. Miao and W. Lee, "Wavelet neural networks for function learning," *IEEE Trans. Signal Processing*, vol. 43, No. 6, pp. 1485-1497, 1995.

[Zhang92]

Zhang, Q. and A. Beneviste, "Wavelet Networks," *IEEE Trans. Neural Networks*. vol. 3, No. 6, pp. 889-898, Nov. 1992.

[Zip78]

Zipp, P., "Effect of electrode parameters on the bandwidth of the surface EMG power density spectrum," *Medical and Biological Eng. Comput.*, **16**, 537-541, 1978.

Appendix A

The Backpropagation Algorithm

By far the most widely used supervised learning method is the backpropagation algorithm, which is a gradient search extended to multiple layers, made possible by a differentiable nonlinearity in each node. Before deriving the learning algorithm, the following notation is introduced:

| Notation | Meaning |
|-------------------------|--|
| $u_{\ell,j}$ | output of the j^{th} node in layer ℓ |
| $w_{\ell,j,i}$ | weight which connects the i^{th} node in layer $\ell - 1$ to the j^{th} node in layer ℓ |
| $\mathbf{x}^{(p)}$ | p^{th} training sample |
| $u_{0,i}$ | i^{th} component of the input vector |
| $d_j(\mathbf{x}^{(p)})$ | desired response of the j^{th} output node for the p^{th} training sample |
| N_ℓ | number of nodes in layer ℓ |
| L | number of layers |
| P | number of training patterns |

For notational convenience, let the 0^{th} layer of the network hold the input vector components: $u_{0,j} = \mathbf{x}_j$. Also, in order to account for the bias weights, let the 0^{th} component of the input vector to each layer be equal to 1; that is, $u_{\ell,0} = 1$. This, in turn, implies that $w_{\ell,j,0}$ are the bias weights. With this notation, the output of a node in layer ℓ is given by:

$$u_{\ell,j} = f(s_{\ell,j}) \quad (\text{A.1})$$

where

$$s_{\ell,j} = \sum_{i=0}^{N_{\ell-1}} w_{\ell,j,i} u_{\ell-1,i}. \quad (\text{A.2})$$

Assume for this derivation that $f(\bullet)$ is the sigmoid nonlinearity. The derivative of this function is

$$f'(\alpha) = \frac{df(\alpha)}{d\alpha} = f(\alpha)(1-f(\alpha)). \quad (\text{A.3})$$

The backpropagation algorithm uses a gradient search technique to find the network weights that minimizes a *sum of squared errors* criterion function:

$$J(\mathbf{w}) = \sum_{p=1}^P J^{(p)}(\mathbf{w}) \quad (\text{A.4})$$

where \mathbf{w} is a vector representing the network weights, and $J^{(p)}(\mathbf{w})$ is the total squared error for the p^{th} pattern:

$$J^{(p)}(\mathbf{w}) = \frac{1}{2} \sum_{q=1}^{N_L} [e_q(\mathbf{x}^{(p)})]^T e_q(\mathbf{x}^{(p)}), \quad (\text{A.5})$$

where $e_q(\mathbf{x}^{(p)}) = d_q(\mathbf{x}^{(p)}) - u_{L,q}(\mathbf{x}^{(p)})$ is the output error at each node. Here, N_L is the number of nodes in the output layer (the number of classes). This sum squared error function defines an error performance surface that is equal in dimension to the number of network weights. The gradient search algorithm seeks a minimum of this surface, hopefully global rather than local. The weights of the network are determined iteratively according to:

$$\begin{aligned} w_{\ell,j,i}(k+1) &= w_{\ell,j,i}(k) - \mu \left. \frac{\partial J(\mathbf{w})}{\partial w_{\ell,j,i}} \right|_{\mathbf{w}(k)} \\ &= w_{\ell,j,i}(k) - \mu \sum_{p=1}^K \left. \frac{\partial J^{(p)}(\mathbf{w})}{\partial w_{\ell,j,i}} \right|_{\mathbf{w}(k)} \end{aligned} \quad (\text{A.6})$$

where μ is a positive constant called the *learning rate*. To implement this algorithm we must develop an expression for the partial derivative of $J^{(p)}$ with

respect to each weight in the network. For an arbitrary weight in layer ℓ this can be written using the Chain Rule:

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial w_{\ell,j,i}} = \frac{\partial J^{(p)}(\mathbf{w})}{\partial u_{\ell,j}} \frac{\partial u_{\ell,j}}{\partial w_{\ell,j,i}} \quad (\text{A.7})$$

or

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial w_{\ell,j,i}} = \frac{\partial J^{(p)}(\mathbf{w})}{\partial s_{\ell,j}} \frac{\partial u_{\ell,j}}{\partial s_{\ell,j}} \frac{\partial s_{\ell,j}}{\partial w_{\ell,j,i}}. \quad (\text{A.8})$$

This reduces to

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial w_{\ell,j,i}} = \delta_{\ell,j} u_{\ell-1,j}, \quad (\text{A.9})$$

where

$$\delta_{\ell,j} \equiv \frac{\partial J^{(p)}(\mathbf{w})}{\partial s_{\ell,j}}. \quad (\text{A.10})$$

Thus, we have a weight update equation:

$$w_{\ell,j,i}(k+1) = w_{\ell,j,i}(k) - \mu \delta_{\ell,j} u_{\ell-1,j}. \quad (\text{A.11})$$

All that is left is to compute the δ 's corresponding to each layer. At the output layer L , we have the boundary condition:

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial u_{L,j}} = u_{L,j}(\mathbf{x}^{(p)}) - d_j(\mathbf{x}^{(p)}) \quad (\text{A.12})$$

such that

$$\delta_{L,j} = -e_j(\mathbf{x}^{(p)}) u_{L,j} (1 - u_{L,j}). \quad (\text{A.13})$$

For all other layers we have

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial u_{\ell,j}} = \sum_{m=1}^{N_{\ell+1}} \frac{\partial J^{(p)}(\mathbf{w})}{\partial u_{\ell+1,m}} u_{\ell+1,m} (1 - u_{\ell+1,m}) w_{\ell+1,m,j} \quad (\text{A.14})$$

or

$$\frac{\partial J^{(p)}(\mathbf{w})}{\partial u_{t,j}} = \sum_{m=1}^{N_{t+1}} \delta_{t+1,j} w_{t+1,m,j} \quad (\text{A.15})$$

yielding

$$\delta_{t,j} = u_{t,j}(1-u_{t,j}) \sum_{m=1}^{N_{t+1}} \delta_{t+1,j} w_{t+1,m,j}. \quad (\text{A.16})$$

When training the network, the weights are typically initialized to small random values. This starts the search in a relatively “safe” position. Each input/output pair in the training set \mathfrak{I} is presented to the network, and the network output is computed in the *feedforward* direction. The error at the output $e_r(\mathbf{x}^{(p)})$ is computed, and the weights in the L^h layer are updated using Equations (A.11) and (A.13). These updates (the δ 's) are then *backpropagated* through the hidden layers so that the weight updates in the hidden layer nodes may be calculated using Equations (A.11) and (A.16). The next input/output pair is presented, and the process repeats using the previously updated weights. If the weights are updated at every pattern presentation, the learning is termed *patternwise*. Alternatively, the errors can be averaged over many patterns (perhaps through an entire cycle of the training set); this is often called *epochwise* learning.

Appendix B

Quadratic Time-Frequency Representations

The quadratic structure of a TFR is intuitively reasonable if one wants to interpret the time-frequency response as an energy distribution (or “instantaneous power spectrum”), since energy is a quadratic signal representation. As such, an “energetic” TFR $T_x(t, f)$ seeks to incorporate the concepts of *instantaneous power* and the *spectral energy density*, which are:

$$p_x(t) = |x(t)|^2 = \int_f T_x(t, f) df \quad (\text{B.1})$$

and

$$P_x(f) = |X(f)|^2 = \int_t T_x(t, f) dt \quad (\text{B.2})$$

respectively. The most fundamental of all quadratic representations is the Wigner-Ville distribution (WVD) [Wigner71][Cohen89]:

$$W_{x,y}(t, f) \doteq \int x\left(t + \frac{\tau}{2}\right) y\left(t - \frac{\tau}{2}\right) e^{-j2\pi ft} d\tau \quad (\text{B.3})$$

or, equivalently:

$$W_{x,y}(t, f) \doteq \int X\left(f + \frac{v}{2}\right) Y\left(f - \frac{v}{2}\right) e^{-j2\pi vt} dv, \quad (\text{B.4})$$

which is the cross-density between signals $x(t)$ and $y(t)$. The auto-WVD may be loosely interpreted as a two-dimensional distribution of signal energy over the time-frequency plane. The auto-WVD is always real-valued, and it preserves time

shifts and frequency shifts of the signal. The strength of the WVD is that the time and frequency concentration of the signal are preserved exactly. This is different from the spectrogram and the scalogram which generally introduce some broadening with respect to time and frequency¹. Figure B.1 demonstrates the TFR of a sinusoid, localized in time by a Gaussian window. The TFR is represented as a spectrogram (using a Hamming window, sliding across the time axis one sample at a time), a scalogram (a continuous wavelet transform using a Morlet wavelet, computed at integer scales of 2 to 15), and a WVD. All TFRs have been plotted on the same time and frequency scale. Clearly, the WVD is the most resolved representation here.

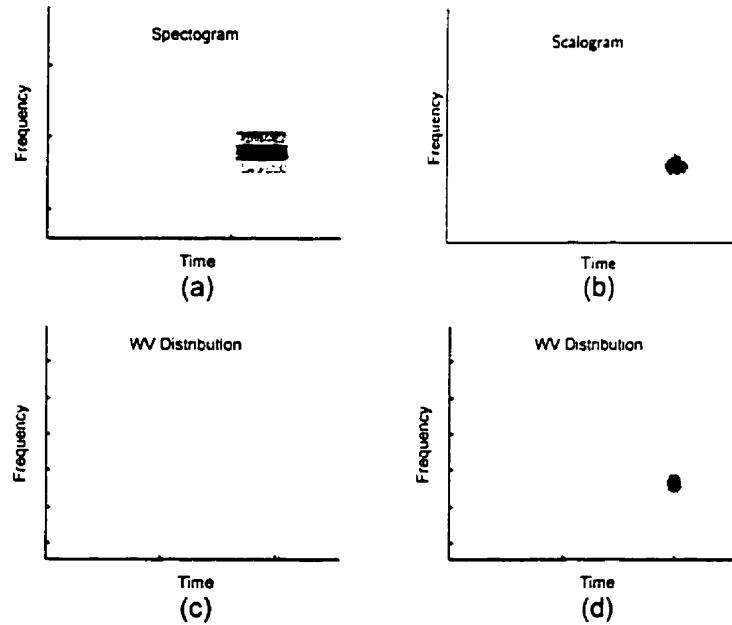


Figure B.1 - The TFR of a time-localized sinusoid. Both the spectrogram (a) and the scalogram (b) exhibit broadening due to their respective analysis windows. The WVD computed from the real-valued signal (c) demonstrates perfect time-frequency resolution, but possesses cross terms due to interference between positive and negative frequency components. Using the analytic signal to compute the WVD (d) substantially reduces these interference terms.

A major issue, however, in the application of quadratic TFRs, is the existence of *cross-terms* in the distribution. In the WVD of Figure B.1(c), there are obvious

¹ The spectrogram and the scalogram are quadratic time-frequency representations, since they are the squared

cross terms present at the same temporal location as the sinusoid, but centered about the zero frequency axis. This is due to interference between the positive and negative frequency components in the double-sided spectrum of the real-valued signal. It follows, then that it is almost imperative that the *analytic* equivalent of the real signal be used in a WVD analysis. The analytical signal is computed from the real signal by taking the Fourier transform, setting the negative frequency components to zero, and then computing the inverse Fourier transform (this operation is known as the *Hilbert transform*). The WVD TFR of the analytic localized sinusoid is shown in Figure B.1Figure (d); the interference terms have been eliminated due to the absence of negative frequency components. It will be assumed, for the remainder of the discussion here, that the WVD is computed from an analytic signal.

Compensation for cross terms due to interference between positive and negative frequency components is a trivial problem, however, compared to that presented by interference between time-frequency components of a multi-component signal. Consider a signal that is composed of two components $x(t) = c_1x_1(t) + c_2x_2(t)$. The quadratic TFR, then, must be

$$_z(t,f) = |c_1|^2 T_{x_1}(t,f) + |c_2|^2 T_{x_2}(t,f) + c_1 c_2^* T_{x_1,x_2}(t,f) + c_2 c_1^* T_{x_2,x_1}(t,f). \quad (\text{B.5})$$

The cross-terms T_{x_1,x_2} and T_{x_2,x_1} can be extremely problematic if they obscure or distort the interpretation of the auto-terms. For a signal with P components, the TFR will comprise P signal terms and $\binom{P}{2}$ interference terms. The number of interference terms grows quadratically with the number of components, often

magnitude of the STFT and the WT, which are linear.

making visual analysis of multicomponent signals difficult. This effect is demonstrated in Figure B.2, which depicts the TFR of a signal consisting of three sinusoids independently localized in time and in frequency. In this figure, the TFR generated by a spectrogram, a scalogram and a WVD have been shown on identical time-frequency axes. The WVD in Figure B.2(c) shows exact localization of the sinusoids (the response at the vertices of the “triangle”), but it exhibits severe interference between components. The spectrogram and scalogram show little cross-term interference, but at the expense of compromised resolution in time, frequency or both.

Consider any two terms $x_1(t)$ and $x_2(t)$, located at (t_1, f_1) and (t_2, f_2) , respectively. The geometry of the interference term between $x_1(t)$ and $x_2(t)$ is such that it resides at $\left(\frac{t_1 + t_2}{2}, \frac{f_1 + f_2}{2}\right)$, oscillates with respect to time at $|f_1 - f_2|^{-1}$ and with respect to frequency at $|t_1 - t_2|^{-1}$ [Hlawatsch92a]. Clearly, the pairwise terms are a significant entity in the WVD TFR of the multi-component signal shown here.

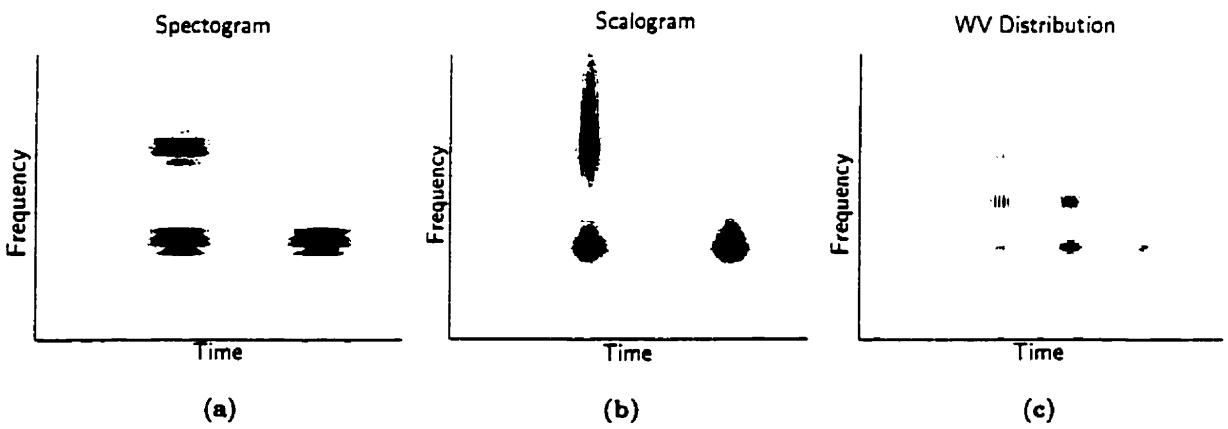


Figure B.2 – The TFR of a multicomponent signal. Both the spectrogram (a) and the scalogram (b) exhibit fair time-frequency resolution, and a mild degree of cross-component interference. The WVD (c) demonstrates superior localization, but severe cross-component interference.

Efforts have been made to eliminate WVD cross-terms by introducing a *smoothing function*, the purpose of which is to attenuate the oscillatory interference terms. Attenuation of the cross-terms comes with a tradeoff, however; the smoothing function must necessarily compromise the time-frequency resolution since it effectively performs a two-dimensional lowpass filtering operation. Indeed, an entire class of quadratic TFRs known as Cohen's class may be expressed as a two-dimensional filtered WVD [Cohen89]:

$$_x(t,f) = \iint_{\tau v} \Psi_T(t-\tau, f-v) W_x(\tau, v) d\tau dv \quad (\text{B.6})$$

where $W_x(t,f)$ is the WVD and $\Psi_T(t,f)$ is the smoothing function (or *kernel*) to obtain the desired TFR, $T_x(t,f)$. The characteristic property of Cohen's class is that the transforms are shift-invariant: time shifts in the signal are preserved in the TFR. Through the selection of the proper kernel, it is easily shown that the spectrogram is a member of Cohen's class of distributions. The spectrogram's kernel substantially smoothes the WVD, eliminating most cross terms, except those between components that are very close together. This comes at the expense of reduced time-frequency resolution. Other examples of TFRs in Cohen's class are described in [Hlawatsch92b]². Each member of Cohen's class attempts to resolve the interference-resolution tradeoff with a different approach; the success of each depends strongly on the nature of the data and the application.

² The scalogram is a member of a related class of quadratic TFRs known as "Affine Distributions", which include both shift invariance and scale invariance as characteristic traits.

Appendix C

Time-Frequency Plane Tiling of Wavelet and Cosine Packet Transforms

A complete wavelet packet or cosine packet binary tree contains many more subspaces than are needed to partition the frequency (wavelet packets) or time (cosine packets) axis. Each subspace has its own time-frequency localization characteristic. The selection of the best basis essentially specifies a single, complete partition *via* a subset of these subspaces, providing one orthonormal basis of $L^2(\mathbb{R})$.

Wavelet Packets

Consider the localization of a given wavelet packet in the time-frequency domain. As depicted in Figure C.1, the bounds of the information cell are described by the vertical distance to the bottom of the cell f , the height of the box Δf , the horizontal distance to the left side of the cell t , and the width of the box Δt .

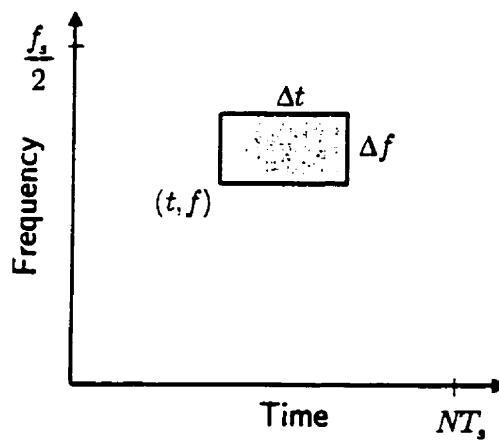


Figure C.1 – An information cell in the time-frequency plane.

The scale, frequency, and position indices (j, k, n) of wavelet packets define the bounds of the information cell. Consider a signal of length $N = 2^{n_0}$ (where n_0 is a positive integer) sampled at f_s Hz, and assume that a wavelet packet analysis is carried out to the lowest possible level $j = n_0$. Then, we have

$$0 \leq j \leq n_0,$$

$$0 \leq k \leq 2^j, \text{ and}$$

$$0 \leq n \leq 2^{n_0-j}.$$

This means that,

$$\Delta t = 2^j \cdot T_s,$$

$$\Delta f = 2^{-j} \cdot \frac{f_s}{2}$$

as with wavelet transform time-frequency localization, as described in Section 3.2.5. With the wavelet transform, the distance f to the lower bound of a cell is given by $f = \Delta f \cdot q$, where $q = 1$ if the subband consists of *detail* coefficients, and $q = 0$ if it is an *approximation* subband. With the wavelet packet transform, q may take many more values since, for a given scale j , the frequency k has the range $0 \leq k \leq 2^j$. As mentioned in Chapter 3, the frequency parameter k is *Paley-ordered*: the frequency range of the corresponding wavelet packet does not increase monotonically with k . To correct this, the frequency parameter must be corrected by *inverse Gray-code permutation* [Simon95], which will be denoted here by the operator GC^{-1} . Thus,

$$f = \Delta f \cdot GC^{-1}(k).$$

The approximate horizontal position is given trivially by

$$t = n \cdot \Delta t.$$

The horizontal positioning is actually only *exact* for Gaussian wavelets, but it is very close for most QMFs. The precise shift to correct the horizontal position for a given QMF is given in [Hess-Nielsen96].

Cosine Packets

For the purpose of determining the time-frequency tiling, it makes no difference if the chosen basis consists of sines or cosines. One is obtained from the other simply by reversing the direction of time: the geometry of the cells is unaffected. Recall that the indexing convention for cosine packets is (j, k, n) , where the triplet corresponds to scale, position (window index) and frequency, respectively. The bounds on these indices are

$$0 \leq j \leq n_0,$$

$$0 \leq k \leq 2^j, \text{ and}$$

$$0 \leq n \leq 2^{n_0-j}.$$

Now, for every level of decomposition, the width of the temporal interval is halved, so that:

$$\Delta t = \frac{NT_s}{2^j} = 2^{n_0-j} \cdot T_s,$$

and from the Heisenberg uncertainty principle:

$$\Delta f = 2^{j-n_0} \cdot \frac{f_s}{2}.$$

The horizontal location is given by

$$t = k \cdot \Delta t = k \cdot 2^{n_0-j} \cdot T_s$$

and the vertical position by

$$f = n \cdot \Delta f = n \cdot 2^{j-n_0} \cdot \frac{f_s}{2}.$$

There is no Gray coding or position correction necessary to properly interpret the time-frequency localization of the cosine packet bases, as with the wavelet packet case.

Appendix D

The Lack of Shift Invariance of the Wavelet Transform

The wavelet transform is a form of subband coding, in which a signal is decomposed into a set of subbands, with the intention that the information within each subband may be processed more or less independently of the others. The discrete wavelet transform is a critically sampled transform, i.e. the number of samples in the transform is equal to the number of samples in the signal. This is possible by the downsampling (by a factor of 2) operation intrinsic to the quadrature mirror filters (QMFs) used in the decomposition and reconstruction algorithms. The QMFs consist of a lowpass filter (h), a highpass filter (g), and their corresponding downsample-by-two operations. This is illustrated in Figure D.1, in which a signal \mathbf{x} is subject to a wavelet decomposition.

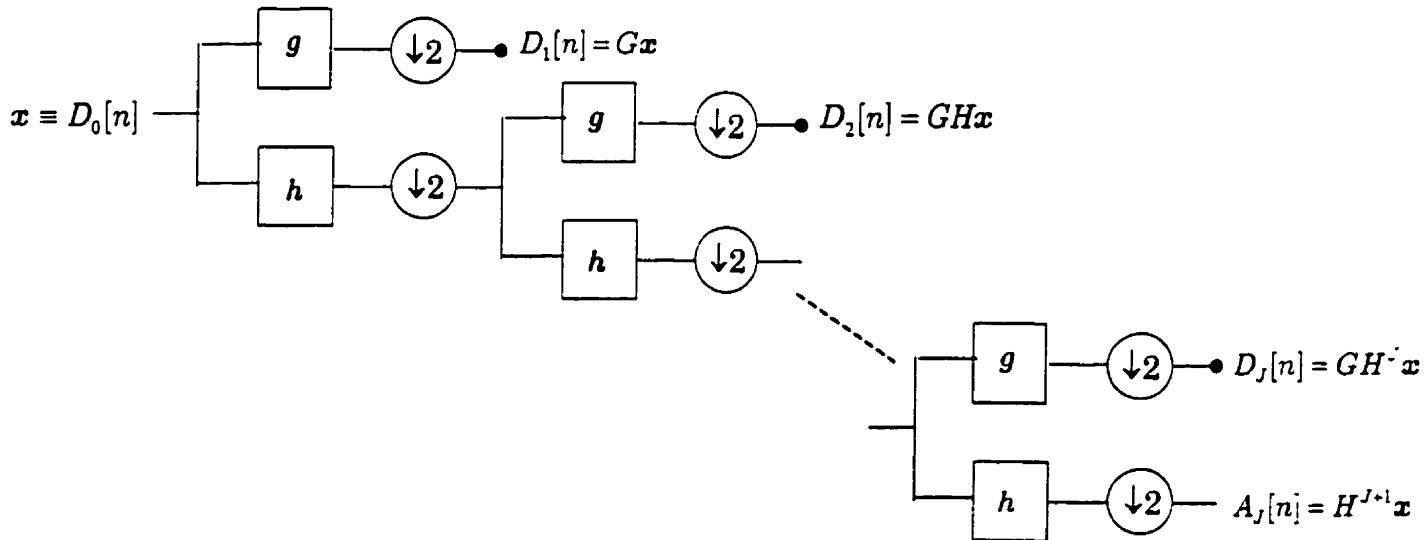


Figure D.1 - The subband coding analogy of the DWT.

This downsampling by two is in direct violation of the Nyquist criterion, and therefore aliasing results during wavelet decomposition. No information is lost, however, as the quadrature mirror filters are orthogonal subband transforms, and ensure that the aliasing errors from all of the subbands cancel when the bands are recombined during reconstruction.

This aliasing phenomenon is illustrated in Figures D.2 (a) and (b).

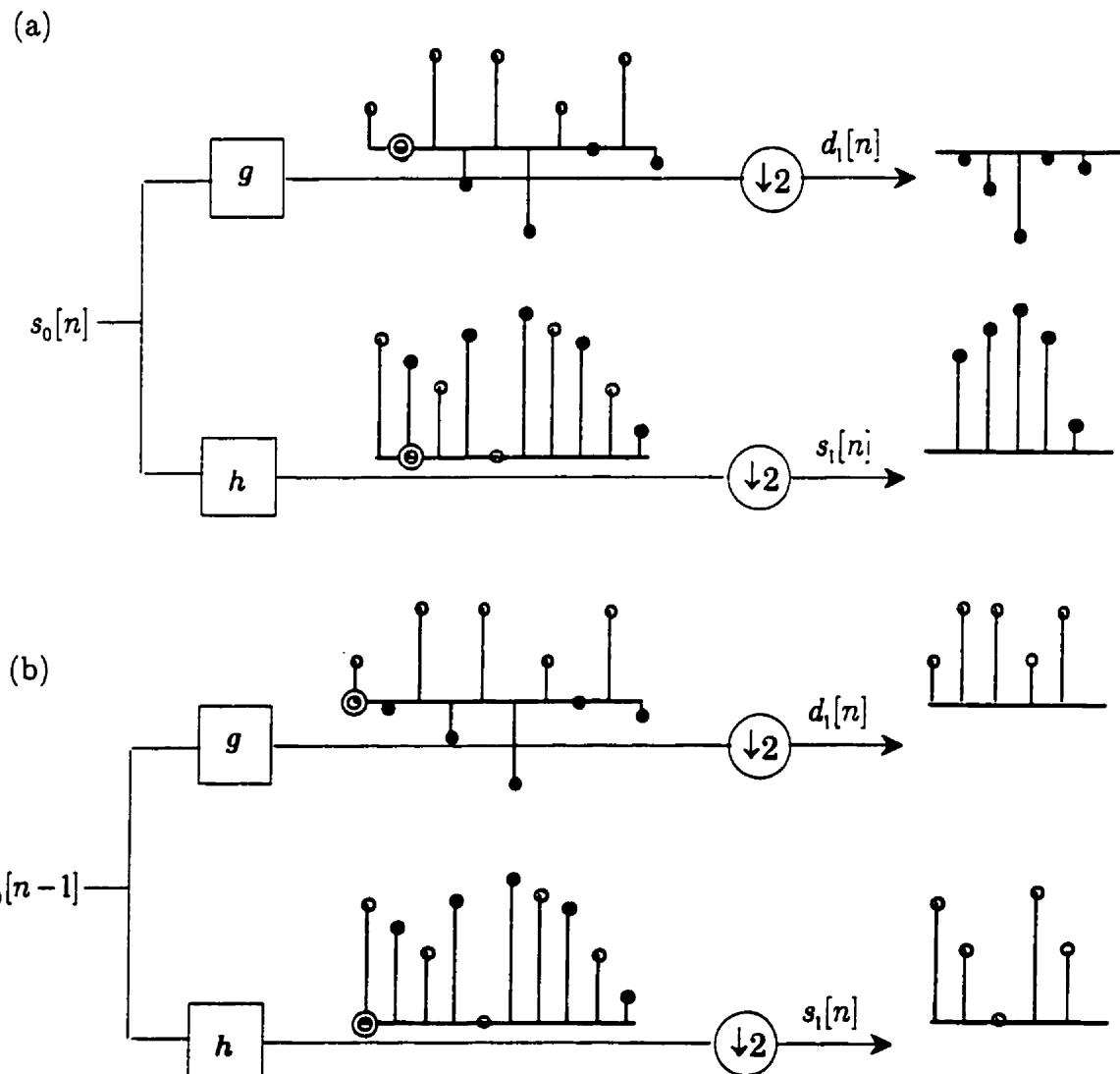


Figure D.2 – Output of the filter bank at the first stage for different inputs: (a) original input; (b) the original input shifted by one.

In Figure (a), a signal $s_0[n]$ is presented to the QMFs to produce the detail coefficients $d_1[n]$ and smooth coefficients $s_1[n]$. The decimated subband signals $d_1[n]$ and $s_1[n]$ are the same for zero shift and all even shifts. In Figure (b), the input signal has been shifted by a single sample but clearly, the decimated subband signals are very different than in (a). This process continues at each stage, propagating the aliases through each level of the decomposition. Therefore, the wavelet coefficients can be very different than those of the original (unshifted) input signal.

We may say that a transform is *translation invariant* if, for a simple shift in the input signal, the transform coefficients experience the same simple shift. Due to the aliasing that occurs in a discrete wavelet decomposition, it is clear that the transform is not translation invariant. Figure D.3 shows the wavelet coefficients of a transient MES burst of elbow flexion at levels 1 through 7 of decomposition. Figure (a) shows the coefficients of the original signal, and Figure (b) shows the coefficients of the signal shifted right by one sample.

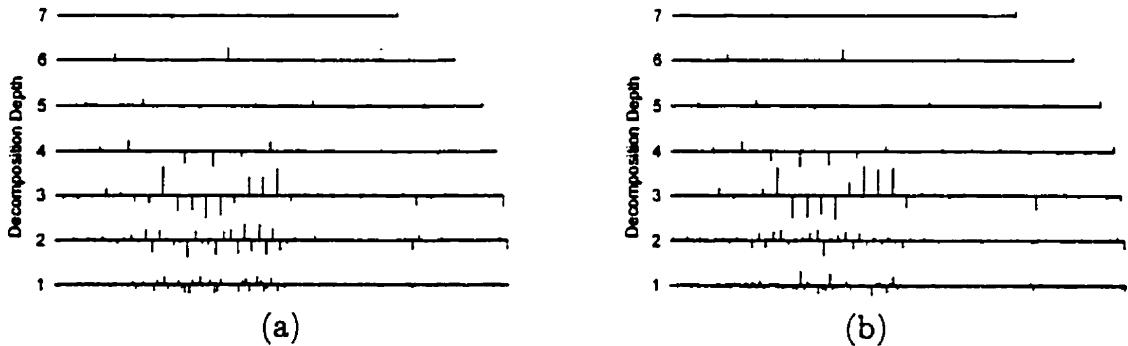


Figure D.3 – The wavelet coefficients of a burst of biceps activity accompanying elbow flexion:
(a) the coefficients of the original signal; (b) the coefficients of the signal shifted right by one sample.

Clearly there is a marked difference in the coefficients at levels one, two and three, which correspond to the high frequency bands. The effect of the aliasing diminishes as it propagates down each level of the decomposition. Figure D.4

depicts the time-frequency response corresponding to each of these decompositions. Again, the aliasing has its most significant effect in the high frequency subbands.

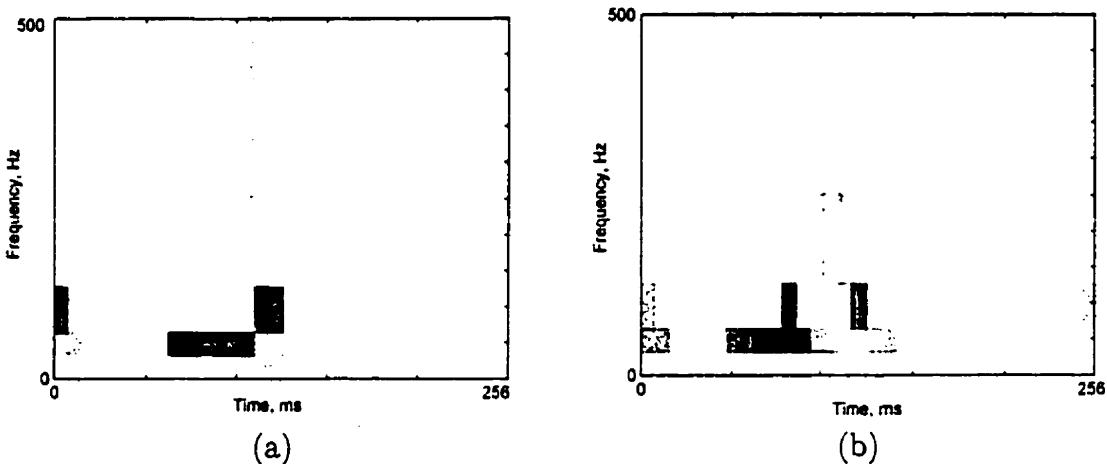


Figure D.4 – The time-frequency response corresponding to wavelet coefficients of a burst of biceps activity accompanying elbow flexion: (a) the TFR of the original signal; (b) the TFR of the signal shifted right by one sample.

Clearly, this has an important effect upon a feature set derived from the wavelet coefficients. The lack of shift invariance will contribute to a nonlinear modification of the wavelet coefficients if the transient MES patterns are subject to random offset due to the means by which they are detected using an amplitude threshold. This will generate a dispersion in the time-frequency plane that is much more complex than simple temporal shifts. This will increase the degree of intra-class variance, and presumably, degrade the generalization capability of the feature set.

Characterizing the Effects of Shift

It is important to know how the effects of WT coefficient modification due to shift will affect the performance of CS and PCA-reduced feature sets. An arbitrary waveform was translated by a series of shifts (-20,-19,...,0,...,19,20). Each WT coefficient was modified in such a manner that no obvious relationship seemed to exist with the degree of shift. Indeed, the change in the WT coefficients appeared to be randomly distributed. This suggests that the modification of WT due to translation may be modeled as *noise*.

If the WT coefficient modification can be modeled as noise, then this has some important implications when using PCA dimensionality reduction. PCA has the ability to “ignore” uncorrelated noise, relegating it to the lesser principal components. To corroborate this proposition a simple dataset was constructed. Four template patterns of transient MES (one from each class of an actual dataset) were replicated by shifting the template by a random amount. The shift was uniformly distributed over a range $[-L, L]$, where L is defined to be a maximum shift limit. A training, validation and test set was generated in this manner. Since the patterns within each class are identical except for the shift, any dispersion of the WT coefficients is due to shift.

The effect of this shift upon the CS-reduced WT features was to introduce this “noise” due to WT coefficient modification throughout all features. A shift limit of eight samples was introduced into the artificial dataset. Figure D.5 shows a scatterplot of the top eight WT coefficients chosen by CS.

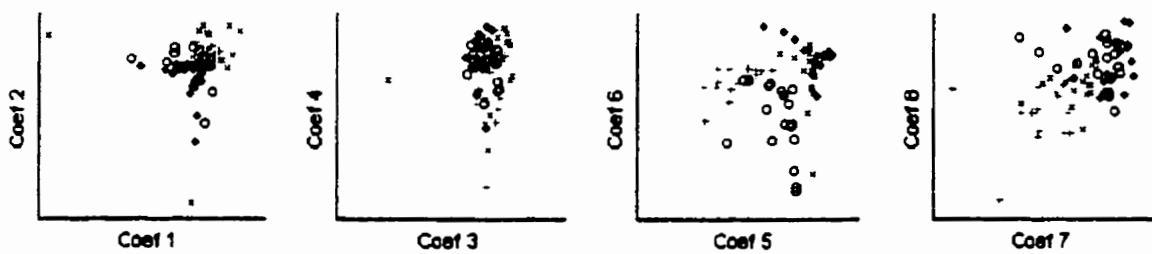


Figure D.5 – The effect of shift-induced dispersion in the CS-reduced WT feature set. The coefficients are presented in pairs, according to their rank.

Clearly, the dispersion is sufficient to obscure the separability amongst the classes. The advantage of PCA however, is that the signal and “noise” subspaces are segregated. Figures D.6 (a)-(c) show scatterplots of the top eight PCA-reduced WT coefficients, corresponding to datasets with a shift limit of four, six, and eight samples, respectively.

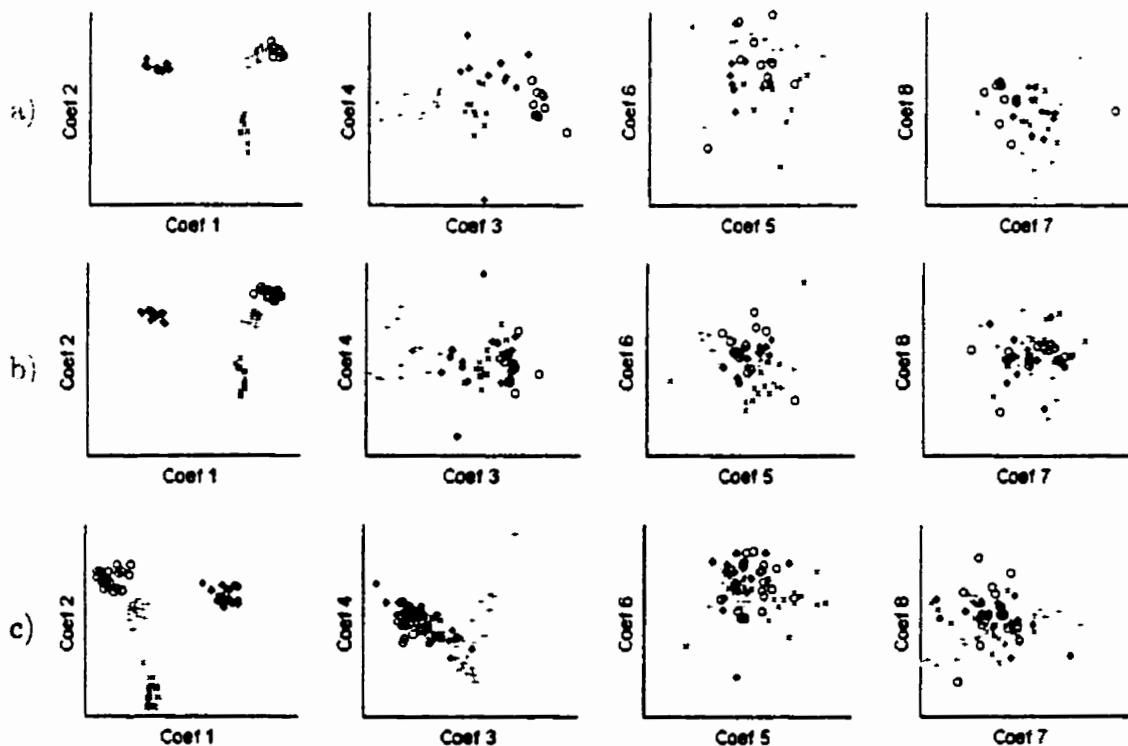


Figure D.6 – The effect of shift-induced dispersion in the PCA-reduced WT feature set. The datasets comprise a shift limit of (a) four, (b) six, and (c) eight samples.

In these scatterplots, we can see that the first two PCA features are those that are providing the most discriminant information. In Figure (a) with a maximum shift of four, the third and fourth PCA features are reasonably separable as well. The lower-ranking PCA features show no discriminant ability at all, as they are completely dedicated to modeling the noise. As the shift limit increases to six and eight, the WT modification noise begins to invade the third and fourth PCA features, eroding their discriminant capacity. Although PCA tends to relegate the noise to the lesser PCA features, it does so only partially, allowing some noise to infiltrate the leading PCA features. Observation of datasets corrupted by incasing levels of actual random noise show effects that are strikingly similar to those in Figure D.6. This corroborates the presumption of the noise-like characteristic of the effects of shift.

This effect of WT modification noise upon the PCA features is perhaps seen more clearly in a measure of dispersion. For each of the first three PCA features, the WT noise was determined (the difference between the shifted WT coefficient and its unshifted value). Next, the standard error¹ of this noise was computed, yielding a measure of dispersion. This dispersion is shown at various shift levels in Figure D.7.

¹ Standard error is defined here as the standard deviation divided by the mean.

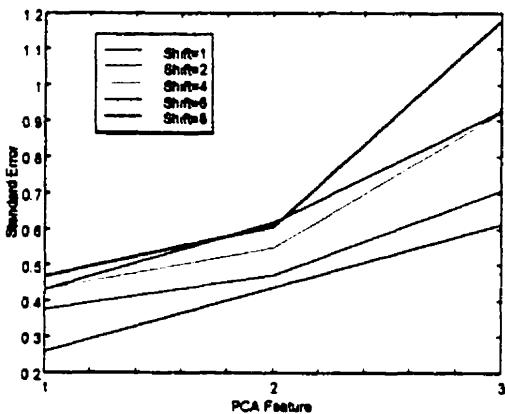


Figure D.7 – The standard error of the dispersion of the first three PCA features. The maximum shift in the artificial dataset comprises the range 1,2,4,6,8.

As expected, increasing values of shift level increase the standard error of the PCA features, which implies greater dispersion. Although the third PCA feature is affected by dispersion to the greatest degree, the first two PCA features are also susceptible to dispersion. Again, this demonstrates that PCA can only partially remove the effects of noise from the leading principal components.

Although this artificial dataset has allowed the effects of WT modification to be examined in isolation, it is important to know how shift affects the WT coefficients in a real dataset. The two channel data from an arbitrary subject was subject to the same range of maximum shift as above. Figure D.8 shows the standard error of the first 20 PCA features in the training set and test set. Twenty PCA features were shown here (as opposed to only three for the artificial dataset) because the problem is more complex. When classifying the unshifted dataset, roughly ten PCA features are required to provide good discrimination.

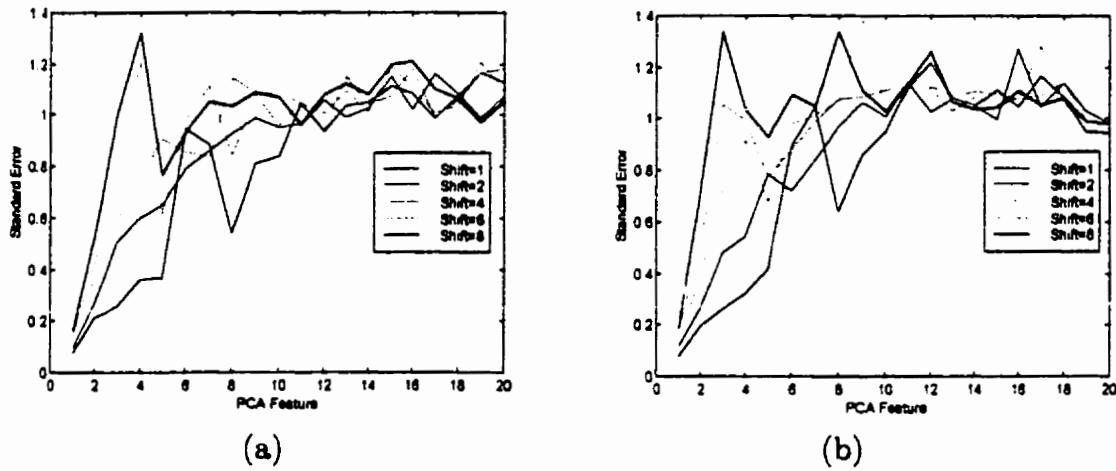


Figure D.8 – The effect of shift upon the standard error of the PCA coefficients when using a real dataset of two channel MES. Figure (a) shows dispersion in the training set, and (b) in the test set.

The effect of shift can be seen to affect those PCA features required to discriminate the data. The dispersion of PCA features 1-10 increases with the level of shift. The same effect is present in both the training set and the test set, implying that PCA's ability to "reject noise", which is learned from training set, generalizes well to the test set. Correspondingly, the test set classification performance degrades as noise due to WT shift infiltrates the leading PCA coefficients. This degradation is not as dramatic however, as that which affects the CS coefficients, as shown in Figure D.9.

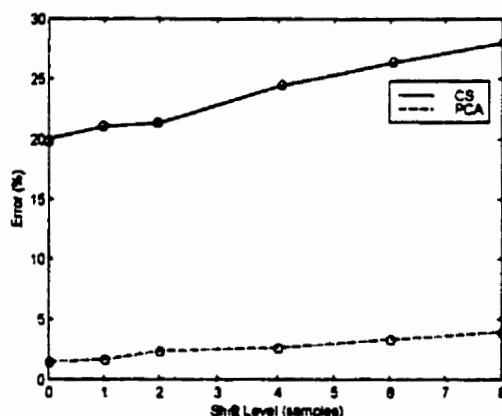


Figure D.9 – The effect of shift upon the classification performance of a dataset of real two channel MES. The performance of CS and PCA-reduced WT feature sets are shown at various levels of shift.

Therefore, it may be concluded that the effect of temporal translation is to introduce what may be considered random noise into the WT coefficients. This noise directly influences the coefficients in feature subset selection schemes such as CS, and severely degrades separability. PCA has the ability to reject this noise from the leading principal components, but only to a certain degree. Shift-induced noise in the WT coefficients will begin to impair the classification performance of a real dataset if the shift is sufficiently pronounced. In the example given here, a shift level of eight inflates the error of the PCA-reduced feature set from 1.5% to 4%.