


This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (<https://dspace.lboro.ac.uk/>) under the following Creative Commons Licence conditions.




CC creative commons
COMMONS DEED


Attribution-NonCommercial-NoDerivs 2.5


You are free:

- to copy, distribute, display, and perform the work

Under the following conditions:

 **Attribution.** You must attribute the work in the manner specified by the author or licensor.


 **Noncommercial.** You may not use this work for commercial purposes.

 **No Derivative Works.** You may not alter, transform, or build upon this work.

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

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

This is a human-readable summary of the [Legal Code \(the full license\)](#).

[Disclaimer](#) 

For the full text of this licence, please go to:
<http://creativecommons.org/licenses/by-nc-nd/2.5/>

BLOSC no: - DX 218306



Pilkington Library

Author/Filing Title LONG

Vol. No. Class Mark T

**Please note that fines are charged on ALL
overdue items.**

DISK IN REAR POCKET.
LOAN COPY

FOR REFERENCE ONLY

CHECK FOR DISK

0402153707



Wavelet Methods in Speech Recognition

by


Christopher Long

A Doctoral Thesis submitted in partial fulfilment of the requirements for
the award of

Doctor of Philosophy of Loughborough University

February 1999

© by Christopher Long 1999

 Loughborough University P... ..y
Date Feb 00
Class
Acc No. 040215707

M000 1250 LB

ABSTRACT

In this thesis, novel wavelet techniques are developed to improve parametrization of speech signals prior to classification. It is shown that non-linear operations carried out in the wavelet domain improve the performance of a speech classifier and consistently outperform classical Fourier methods. This is because of the localised nature of the wavelet, which captures correspondingly well-localised time-frequency features within the speech signal. Furthermore, by taking advantage of the approximation ability of wavelets, efficient representation of the non-stationarity inherent in speech can be achieved in a relatively small number of expansion coefficients. This is an attractive option when faced with the so-called 'Curse of Dimensionality' problem of multivariate classifiers such as Linear Discriminant Analysis (LDA) or Artificial Neural Networks (ANNs). Conventional time-frequency analysis methods such as the Discrete Fourier Transform either miss irregular signal structures and transients due to spectral smearing or require a large number of coefficients to represent such characteristics efficiently. Wavelet theory offers an alternative insight in the representation of these types of signals.

As an extension to the standard wavelet transform, adaptive *libraries* of wavelet and cosine *packets* are introduced which increase the flexibility of the transform. This approach is observed to be yet more suitable for the highly variable nature of speech signals in that it results in a time-frequency sampled grid that is well adapted to irregularities and transients. They result in a corresponding reduction in the misclassification rate of the recognition system. However, this is necessarily at the expense of added computing time.

Finally, a framework based on adaptive time-frequency libraries is developed which invokes the final classifier to choose the nature of the resolution for a given classification problem. The classifier then performs dimensionality reduction on the transformed signal by choosing the top few features based on their discriminant

power. This approach is compared and contrasted to an existing discriminant wavelet feature extractor.

The overall conclusions of the thesis are that wavelets and their relatives are capable of extracting useful features for speech classification problems. The use of adaptive wavelet transforms provides the flexibility within which powerful feature extractors can be designed for these types of application.

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor, Dr. Sekherajit Datta for his support, encouragement and enthusiastic discussions. I was very fortunate to have the opportunity to work with him.

Many thanks are also due to past and present members of the Department of Electrical and Electronic Engineering at Loughborough for their friendship and help over the years. Special thanks also to my former room-mates in the Signal Processing Group for their assistance, helpful discussions and friendship.

I am also very grateful to past and present staff and students at Faraday Hall who made the business of study an altogether more enjoyable experience. In particular, I would like to thank Sharon, James, Nick and David with whom I was fortunate enough to share warden-ship of the hall. We had many memorable times together.

I would also like to thank my present friends and colleagues at the Institute of Psychiatry who knew me during the latter stages of the thesis.

Finally, I would also like to give special thanks to my parents without whose support and guidance a higher degree would have never been possible.

Christopher Long

and no, I haven't forgotten you, Poulakimo.

Loughborough University

February 1999

GLOSSARY

ANN.....	Artificial Neural Network
BB.....	Best Basis
CDF.....	Cumulative Distribution Function
CWT.....	Continuous Wavelet Transform
DCT-IV.....	Discrete Cosine Transform (Type 4)
DECM.....	Discriminant Energy Concentration Measure
DWT.....	Discrete (or Dyadic) Wavelet Transform
FFT.....	Fast Fourier Transform
GCI.....	Glottal Closure Instant
HMM.....	Hidden Markov Model
IF.....	Instantaneous Frequency
JBB.....	Joint Best Basis
KLT.....	Karhunen-Loeve Transform
LCT.....	Local Cosine Transform
LDA.....	Linear Discriminant Analysis
LDB.....	Local Discriminant Basis
LRB.....	Local Regression Basis
LPC.....	Linear Predictive Coding
MWT.....	Multiplexed Wavelet Transform
MRA.....	Multi-Resolutional Analysis
MLRB.....	Modified Local Regression Basis
PCA.....	Principal Component Analysis
PDF.....	Probability Density Function
PSWT.....	Pitch Synchronous Wavelet Transform
STFT.....	Short Time Fourier Transform
TFE.....	Time-Frequency Energy
WP.....	Wavelet Packet

NOTATION

Δt	Time width (Impulse response)
Δf	Frequency width (Bandwidth)
$g(t)$	Fourier basis function (continuous case)
$h(t)$	Mother Wavelet (continuous case)
$\psi(k)$	Mother wavelet (discrete case)
$\ f\ $	Norm(3.3.1)
$\langle f, g \rangle$	Inner product(3.2)

Signals

$f(t)$	Continuous time-domain signal
$f[n]$	Discrete signal

Transforms

$\hat{f}(\omega)$	Fourier transform (3.3.1)
$STFT(\tau, \omega)$	Short-time windowed Fourier transform (3.2)
$CWT(\tau, a)$	Continuous wavelet transform (3.3)
$DWT(\tau, a)$	Discrete wavelet transform (3.3.1)

Sets

\mathfrak{R}	The set of all real numbers
Z	The set of all integers

Spaces

$L^2(\mathfrak{R})$	The space of finite energy functions $\int f(t) ^2 dt < +\infty$
$l^2(Z)$	Finite energy discrete signals $\sum_{n=-\infty}^{+\infty} f[n] ^2 < +\infty$

Information cost measures

$H(\mathbf{p})$	Shannon Entropy (4.3)
$l(\mathbf{p})$	L^2 based cost (4.3)
$M(\mathbf{p})$	Logarithm of the energy (4.3)
$I(\mathbf{p})$	Relative Entropy (5.3.1)

TABLE OF CONTENTS

ABSTRACT	i
ACKNOWLEDGEMENTS	iii
GLOSSARY	iv
CHAPTER ONE : INTRODUCTION	1
1.1 HISTORICAL BACKGROUND ON PARAMETRIZATION TECHNIQUES FOR SPEECH RECOGNITION APPLICATIONS.....	2
1.2 THE BEST-BASIS PARADIGM.....	3
1.3 THESIS OBJECTIVES	4
1.4 THESIS OUTLINE	5
1.5 REFERENCES	6
CHAPTER TWO : REVIEW OF WAVELET METHODS IN SPEECH RECOGNITION	8
2.1 INTRODUCTION	8
2.2 OVERVIEW OF THE WAVELET TRANSFORM	10
2.3 REVIEW OF WAVELETS IN SPEECH RECOGNITION.....	11
2.3.1 Best-Basis Selection for Speech Processing.....	13
2.3.2 Pitch Related Analysis.....	13
2.3.3 Wavelets in Speaker Identification and Authentication	15
2.4 THE RIDGE-SKELETON ALGORITHM.....	15
2.5 SUMMARY	18
2.6 REFERENCES	20
CHAPTER THREE : DISCRETE WAVELET AND FOURIER TRANSFORMS FOR PHONEME RECOGNITION	26
3.1 INTRODUCTION	26
3.2 THE SHORT TIME FOURIER TRANSFORM.....	27
3.3 THE WAVELET TRANSFORM.....	30
3.3.1 Some Fundamental Wavelet Definitions.....	32

3.4	RESULTS.....	42
3.5	DISCUSSION.....	47
3.6	SUMMARY	48
3.7	REFERENCES	64

CHAPTER FOUR : THE BEST BASIS ALGORITHM FOR PHONEME

	CLASSIFICATION.....	66
4.1	INTRODUCTION.....	66
4.2	WHY WAVELET/TRIGONOMETRIC PACKETS?	68
	4.2.1 Wavelet Packet Bases.....	69
	4.2.2 Lapped Orthogonal Transforms.....	73
	4.2.3 Translation Invariance	77
4.3	SELECTION OF THE 'BEST- BASIS'	78
4.4	LINEAR AND NON-LINEAR APPROXIMATION IN BASES	83
	4.4.1 The Linear Case.....	83
	4.4.2 The Non - Linear Case.....	85
4.5	RESULTS.....	87
4.6	DISCUSSION.....	88
4.7	SUMMARY	90
4.8	REFERENCES	99

CHAPTER FIVE : DISCRIMINANT WAVELET BASES AND

	APPLICATIONS.....	101
5.1	INTRODUCTION.....	101
5.2	PROBLEM DEFINITION.....	102
5.3	THE LOCAL DISCRIMINANT BASIS ALGORITHM.....	104
	5.3.1 Cost Measures.....	105
	5.3.2 Non-Additive Criteria.....	109
5.4	RESULTS.....	113
5.5	DISCUSSION.....	117
5.6	SUMMARY	119
5.7	REFERENCES	125

CHAPTER SIX : CONCLUSIONS	127
6.1 OVERVIEW	127
6.1.1 Dyadic Wavelet Transform	129
6.1.2 The Best Basis Algorithm.....	130
6.1.3 Discriminant Wavelet Methods	130
6.2 FUTURE WORK	132

APPENDIX : PUBLISHED PAPERS

Chapter One

Introduction

Typical speech recognition systems are composed of several stages, of which two merit particular interest. The *pre-processing* stage where the raw speech data is appropriately conditioned for the subsequent module and the *classifier* which gives a likelihood measure to each speech frame it encounters. The classifier itself is fed by a *parametrization* stage whose purpose is to enhance recognition by simultaneously reducing the complexity and size of the data. This process is sometimes termed *feature extraction*, and aims to find the best subset of parameters for recognition. As Parsons points out in [9] – “an ideal set of features selected for recognition should meet the following criteria:

- a) Vary widely from class to class.
- b) Insensitive to extraneous variables (i.e. text, context, health and emotional state of talker, system transmission characteristics, etc.).
- c) Stable over long periods of time.
- d) Frequently occurring.
- e) Easy to measure.
- f) Not correlated with other features. ”

In this thesis, alternative ways of extracting a set of such features satisfying some or all of the above requirements using wavelet methods will be examined. The performance of the feature set will be measured in terms of the overall error rate of the speech recognition system.

This chapter will first introduce speech parametrization methods and their various characteristics before briefly outlining the thesis itself.

1.1 Historical Background on Parametrization Techniques for Speech Recognition Applications

Approaches such as Linear Predictive Coding (LPC) [8] and the Windowed Fourier Transform [8] are widely used in speech processing. They are generally fast, robust and give rise to easily interpretable feature sets; one can easily extract formant or pitch information from a speech signal by observing the magnitude of its spectra. In the past, techniques used for parametrizing speech have been based upon criterion derived from *a priori* knowledge of the auditory system. The resultant assumptions are usually reasonable; for example, if *mel-scale warping* is applied to FFT (Fast Fourier Transform) coefficients [8], this has the effect of emphasising perceptually meaningful frequencies. A related but differing notion using the same LPC coefficients is known as PLP (Perceptual Linear Prediction) [1]. While the resulting features may contain discriminant information suitable for speech parametrization, one or more of the following issues regarding the assumptions of these methods may come to mind:

- (i) The auditory system (human ear and brain) does not behave like a frequency spectrum. Even if one successfully models the frequency characteristics of the inner ear, little is known about subsequent processing in the brain. Furthermore, within the speech community it has been accepted e.g. [10] that linear models are inappropriate for speech analysis.
- (ii) Fourier analysis is unable to give simultaneous time-frequency localisation of features. Fourier analysis provides an approximation of a given signal by using a weighted sum of building blocks or *basis functions* which are sine or cosine functions. As each term in the expansion has a particular frequency, a Fourier approximation gives information regarding the *frequency* of a signal. However, in its basic form, Fourier analysis gives no information about frequency behaviour over time. Thus it gives good estimates of global signal characteristics but disregards locally changing events. In the context of speech analysis where the signal does not remain stationary over long periods of time, one requires an analysis technique which takes account of the 'short-term' events. One such method is the Short-Time

Fourier Transform which performs local analysis by windowing the signal in question, however this is not without its drawbacks e.g. spectral leakage.

Bearing in mind these issues, one may turn to alternative methods which address one or both of the preceding problems. One such possibility is the *wavelet transform* which has found application in speech recognition to a limited extent but has made a much bigger impact in fields like image compression. It has some relation to Fourier methods in that real world functions are approximated using basis functions but these are wavelets instead of sines and cosines. Families of wavelets can be generated by dilating (i.e. altering the support of the function) *and translating* a single *mother wavelet*. They fundamentally differ from Fourier bases in that they are well-localised in both the time and frequency domains. In fact Daubechies [4] discovered a family of wavelet basis functions that were simultaneously (i) orthonormal, (ii) have compact support (i.e. zero outside a finite interval), and (iii) have a variable degree of smoothness that can be chosen. The wavelet transform (WT) of a signal thus carries information about the variation of frequency with time.

A further advantage of wavelets is that a small number of wavelet coefficients can be used to approximate efficiently the discontinuities or irregularities often found in real world signals. Generally, in speech short term transitory events contain the most information [8] so a more compact, accurate representation is clearly desirable. Fourier and LPC-based methods tend to lose or average such features within the time window. The standard wavelet transform however, overcomes this shortcoming by analysing the signals in a similar way to an octave band filter bank (see Chapter 3 for details), in other words high frequency events are analysed using a window of compact time support whilst low frequency smooth sections, which are generally of longer duration, are analysed through a correspondingly longer analysis window.

1.2 The Best-Basis Paradigm

In recognising some of the difficulties encountered in representing highly variable speech signals to a classifier, one can easily realise the attraction of adaptively tiling the time-frequency (TF) plane. Windowed Fourier methods and wavelet transforms differ in the time-frequency sense; in that the windowed Fourier transform has a fixed window, while the window support of the WT varies logarithmically with frequency. Since the

standard wavelet transform is non-adaptive, the Best-Basis (BB) Paradigm [2] was developed as a generalisation of the wavelet transform which can tile the TF plane according to the information content of the signal, information in this case is in the Shannon-Entropy sense. Thus, the algorithm adaptively selects a set of basis functions suited to a particular problem by minimising some kind of criterion or *cost function*. The type of cost function depends on the final problem. Conceivably this could be compression, regression or classification. In the latter case, one would ideally desire an adaptive wavelet basis which maximises the distance (in some sense) between classes in the training set. Once the best set of subspaces has been chosen, one can perform “Dimensionality Reduction” by sorting the resulting set of coefficients, expanded in this basis according to their importance, again in terms of the final problem. This step has the dual advantage of increasing the robustness of the classifier (if classification is the problem) by reducing the amount of superfluous or redundant information in the speech signal, while increasing the training speed. Furthermore, if a good discriminant set of coefficients is chosen, further analysis by speech scientists or linguists may yield new insight into the characteristics of speech.

The above discussions encompass the aims of this thesis; parametrization of speech signals using wavelets and their relatives to best fulfil the objectives of Parsons [9], outlined above. Since the wavelet transform is conceptually similar to Fourier methods, future discussions and analysis are generally restricted to this analogy.

1.3 Thesis Objectives

Wavelet applications, although well developed in areas such as signal and image compression, have found relatively limited use in feature extraction and discrimination for speech or related tasks. Therefore this thesis has been concerned with the following:

- (i) To quantitatively compare the dyadic wavelet transform with short time Fourier methods in relation to misclassification performance on different speech classification problems. This includes the assessment of the effects of using different types of wavelets and examining whether wavelets preserve the acoustic-phonetic attributes of speech signals.

- (ii) To implement and compare the Best-Basis algorithm which, due to its adaptivity should improve the performance of the final classifier. Also, it is relevant to decide whether the translation-invariance problem of wavelets is important for speech recognition applications.
- (iii) To develop a discriminant wavelet based feature extractor suitable for phoneme classification using the final classifier to choose the best set of features.
- (iv) To implement and compare this approach with existing discriminant schemes.

1.4 Thesis Outline

Chapter 2 briefly reviews the basics of wavelet theory and gives a literature review of where wavelets have found application in speech recognition. Most of this work is largely based on the standard WT, taking advantage of its multiresolutional properties to characterise transient events. Large scale application to different phonetic subclasses is somewhat lacking. Discriminant wavelet bases are also seen to have found limited use.

In Chapter 3, the theory and properties of wavelets are described in detail, also Fourier and wavelet techniques are compared in their application to several common phoneme classification problems.

In particular, the characteristics of the wavelet are noted which enable it to represent irregular signals and transients via their so called 'zooming property'.

Chapter 4 applies the BB Paradigm, which adaptively tiles the time-frequency plane. In the experiments, the final expansion coefficients used for parametrization are obtained only with the aim of representing the most information in the speech signal in the fewest coordinates and so cannot strictly be viewed as feature extraction. Thus the costs used (and compared) are the same as those used in most compression applications. This should result in a more flexible characterisation of subtle acoustic events inherent in speech and the recognition performance is compared with the results obtained in Chapter 3. The findings are analysed and the Best Basis approximation framework is compared with conventional linear techniques in this particular setting. Also considered in this chapter is the well-known translation invariance problem of discrete wavelet techniques. An algorithm known as Cycle Spinning [3], originally designed to remove artefacts in

wavelet denoising, is applied to the speech datasets prior to training and classification. The overall effect on the final misclassification rate is examined.

The BB Paradigm introduces two new types of basis function: wavelet packets and cosine packets, the characteristics of these are described in detail.

In Chapter 5, a novel modification of the BB paradigm, known as the Local Discriminant Basis algorithm (LDB) which was designed for classification in [6], is applied to the speech datasets. There are two main stages in the LDB algorithm:

- (i) Build a class probability model based on the average energies of the member signals in each class.
- (ii) Search all subspaces to find a set which gives the maximum separation or distance between classes.

In practice, the distance used by LDB to provide contrasts among a set of classes is quite different to the criterion which the classifier uses. A modification of the algorithm which uses the same distance measure as the classifier to select the most discriminant basis. This is seen to improve training performance (and prediction in certain cases) compared to LDB.

Chapters 4 & 5 are expanded versions of publications [5], [6], [7] which are included in the Appendix.

1.5 References

- [1] Atal, B.S., Remde, J.R., "A New Model of LPC excitation for producing Natural Sounding Speech at Low Bit Rates," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*.(Paris, France, 1982), pp. 614-617.
- [2] Coifman, R.R. and Wickerhauser M. V. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [3] Coifman, R.R., Donoho, D.L., "Translation-invariant denoising," in *Wavelets and Statistics*, Lecture notes in statistics, ed. Antoniadis and Oppenheim, Springer-Verlag, 1995, pp. 125-150.

- [4] Daubechies, I. "Orthonormal Bases of Compactly Supported Wavelets," *Comm. Pure and Appl. Math.*, 41:909-996, 1988.
- [5] Long, C.J. and Datta, S. "Wavelet Based Feature Extraction for Phoneme Recognition" *Proc. of the 4th Int. Conf. on Spoken Language Processing (ICSLP96)* (Philadelphia PA, USA 1996), vol.1, pp.264-267
- [6] Long, C.J. and Datta, S., "Time-Frequency Dictionaries for Improved Local Feature Extraction," *Proc. of IEE Colloquium on Pattern Recognition*, London, 26 Feb. 1997, pp 9/1 – 9/6.
- [7] Long, C.J. and Datta, S., "Discriminant Wavelet Basis Construction for Speech Recognition." *Proc. 5th Int. Conf. on Spoken Language Processing (ICSLP98)* (Sydney, Australia, 1998) vol.1, pp. 1047-1049.
- [8] O'Shaugnessy, "Speech Communication : Human and Machine," *New York: Addison-Wesley*, 1987.
- [9] Parsons, T. "Voice and Speech Processing." *McGraw-Hill series in electrical engineering, communications and signal processing*. 1987
- [10] Teager, H.M. and Teager, S.M., "A Phenomenological Model for Vowel Production in the Vocal Tract," *In R.G.Daniloff, ed., Speech Sciences: Recent Advances*. San Diego, Calif.: College-Hill Press, pp. 73-109, 1983.

Chapter Two

Review of Wavelet Methods in Speech Recognition

2.1 Introduction

Reliable speech recognition relies upon an efficient representation of the speech signal. Existing parametrization techniques such as Linear Predictive Coding (LPC) and the Windowed (or Short Time) Fourier Transform have many useful applications in speech processing and are well established in terms of theory and application. However, they have their drawbacks. For example, Fourier techniques lack good localisation of potentially relevant speech events due to the global nature of their basis functions. Furthermore, they incorporate assumptions regarding the speech signal such as quasi-stationarity which may affect performance.

The Wavelet Transform is an alternative method that possesses a number of features which make it particularly suited to signal processing applications in general. Put simply, the Wavelet Transform has orthogonal basis functions which, as shown in Figure 2.1, are well-localised in both the time and frequency domain. Also this technique performs Multi-Resolutional Analysis (MRA) on a given signal; that is, it

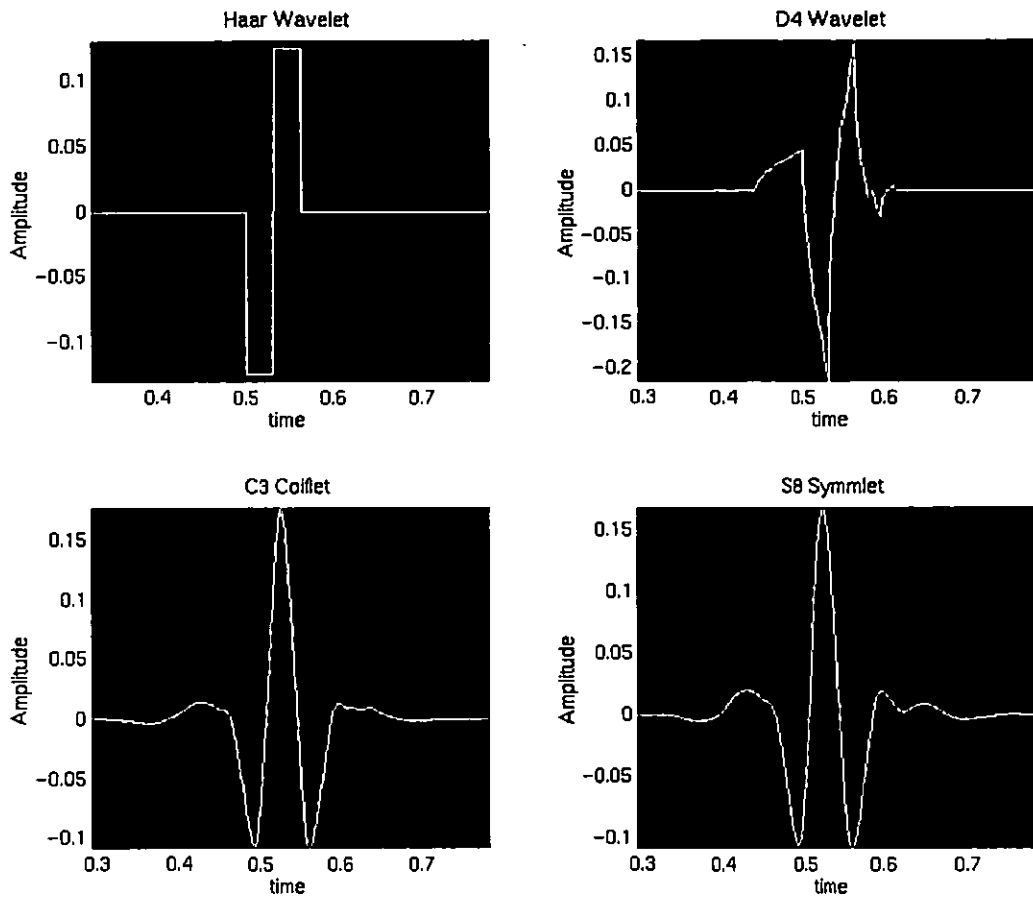


Figure 2.1: Examples of some different wavelet basis functions. D4 is a Daubechies wavelet of order 4, C3 is a Coiflet wavelet with 6 vanishing moments and the S6 is the so-called symmlet wavelet with 8 vanishing moments.

can extract orthogonal sets of information from a signal at several resolutions.

This chapter provides a basic introduction to wavelets and reviews the relevant literature.

2.2 Overview of the Wavelet Transform

The wavelet transform may be thought of as an approximation method which uses the superposition of basis functions (or wavelets) to synthesise an arbitrary function. The relation to Fourier methods is as follows. In the Fourier transform, functions may be approximated by means of a series expansion with sinusoidal terms of differing frequency. This gives rise to the notion of *frequency analysis*.

The human auditory and perceptual system is not sensitive to frequency alone but also to the variation of frequency with time exhibited by most real world signals [9]. A logical extension to the Fourier Transform (FT) which takes this into consideration, is the Short Time Fourier Transform (or STFT). The STFT introduces time dependence into the standard FT whilst retaining the linearity of the operation. In practice, this is achieved by multiplying the time domain signal with a sliding window through which the signal is 'viewed' and the spectra extracted over all times. One limitation is that although global changes in signal behaviour are accounted for, local variations of duration less than the width of the window are averaged or *smear*ed across the time-frame. This is commonly known as *Spectral Leakage*. Since it is known that less stable speech sounds, i.e. those with high transient characteristics carry a significant proportion of speech information, this is clearly an undesirable effect. A further problem with the STFT is that it possesses *constant relative bandwidth* (denoted by Δf , the bandwidth of the window). The windowing function, once chosen, has bandwidth (window duration over frequency) and thus impulse response (window duration over time, denoted by Δt) which remain fixed over all time and frequency. These parameters are dependent upon each other since the response of this window, when drawn on the time-frequency plane, is a rectangle with constant area (as shown in Chapter 3, Figure 3.1). The following inequality is known as the Heisenberg or Uncertainty principle from quantum mechanics which limits how small the area of these tilings can be.

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi} \quad (2.1)$$

From equation (2.1) it is clear that good time resolution can be achieved at the expense of frequency resolution and vice versa. Arbitrarily good resolution in both the time and frequency domains is not possible. Gabor in his original analysis in 1946 [39] suggested the use of Gaussian windows since they met the bound of equation (2.1) with equality.

The wavelet transform on the other hand is an octave band decomposition of the phase plane, which uses the notion of *time-scale analysis*. The translation of frequency in the Fourier transform is replaced by dilation of the basis functions. It is this dilation and contraction that gives rise to the idea of scale. These basis functions, differ from the FT in that they are localised to some extent in both time (or space) and frequency (scale) as opposed to the sine/cosine basis functions of the Fourier transform which are completely global in time and completely localised in frequency. The properties of these approaches is considered in greater detail in Chapter 3.

2.3 Review of Wavelets in Speech Recognition

Wavelets have been used in areas such as formant extraction, [4], [17], [30], [23], signal representation, [5], [8], [14], [31], classification, [3], [13], [20], [21], [34], pitch extraction, references [11], [12], [19], [29], [35], [28], [45], detection of stop consonants, [23], [24] speaker identification and speaker authentication, [22], [27], [38]. They have also been used in speech compression, [43] due mainly to their good approximation ability when applied to irregular signals. Articles of a more general nature relating to the work in this thesis can be found in [1], [4], [2], [6], [7], [10], [26], [33], [37], [43], [25], [43], [16], [32], [15] wherein information on both the fundamentals of wavelet theory and speech recognition as studied in this thesis are discussed. Work on modelling the auditory system has also been considered, [44], [36], [18], [39], [40], the motivation being that improved understanding of the philosophy underlying the human auditory and perceptual system should result in more authentic and therefore superior Automatic Speech Recognition systems (ASR's).

One of the earliest attempts to incorporate wavelets in a speech recognition system was carried out by Davenport et al [8] in which the Wavelet Transform was incorporated directly into the feature extraction stage of the system. In this case, the acoustic-phonetic characterisation of the speech signals (voicing and frication) was obtained from the energy characteristics of the wavelet subband decompositions. The authors used wavelet decomposition prior to classification by a neural network to select frequencies of interest for their particular problem – discrimination between phoneme subclasses using voicing and frication as the cue. They conclude that the most difficult classes to discriminate between were stops, and fricatives.

The concept of decomposing a speech signal into a number of resolutional scales is appealing for it is this which gives the wavelet transform its zooming capability—allowing overall coarse views or, alternatively, giving the option of focussing on fine details of the signal, rather like studying an object with a microscope. However, this property is still subject to certain restrictions. Unlike the STFT where resolution is fixed over the whole time-scale plane wherein the Heisenberg inequality is an issue, the wavelet transform’s resolutional rules are governed by

$$\frac{\Delta f}{f} = c \quad (2.2)$$

where c is a constant and f is the frequency.

Although still requiring a trade-off between time and frequency resolution, equation (2.2) shows that there is good frequency resolution but poor time resolution at low frequencies, and vice-versa at high frequencies. This is demonstrated in the tiling configuration shown by Figure 3.2.

It was these considerations that motivated the concept of *Compound Wavelets*, proposed by Favero [14], in which he discusses the desirability of controlling the time-bandwidth product specifically to enable a more accurate parametrization of speech signals for speech recognition. The benefits are twofold. First, there is an increase in speech recognition performance; second, there is an improvement in computational efficiency due to the reduction in the number of wavelet coefficients required by the process.

D'Alessandro et al [5], [6] uses a more descriptive approach based on linear acoustic theory to model unvoiced speech sounds. Randomly generated wavelets are used as the formant filters in the speech model. The main disadvantage with this method is that once the wavelet coefficients have been calculated, the model cannot adapt to account for variability between speakers. Furthermore, the acoustic parameters of the unvoiced speech source have to be known *a priori*.

2.3.1 Best-Basis Selection for Speech Processing

Wickerhauser in [34], [36] developed the wavelet and cosine packet generalisation from which an application-dependant 'best basis' can be chosen from a library of orthonormal bases. This approach gives a degree of adaptability to the wavelet transform implying its suitability for highly variable signals like speech. Furthermore, if cosine packets are chosen as the basis, they themselves bear a distinct resemblance to bursts of sound, again suggesting their usefulness for the representation of speech signals. The "Best Basis method" was actually developed through the approach of entropy minimisation by Coifman et al [27]. Using this concept, Wesfried et al [5] introduced a speech representation based on the Adapted Local Trigonometric Transform. The window size into which the speech data is partitioned is performed automatically by the BB method; in fact the segmentation depends upon the spectrum it contains. The transitions between windows is thus shown by the authors to be suitable for segmentation into voiced-unvoiced portions. A formant representation is introduced by locating and retaining the centres of mass for the highest-value peaks of the transform. From this, the local spectrum can be seen to represent the formants of the speech signal.

2.3.2 Pitch Related Analysis

Wavelet-based pitch analyses have been proposed by Kadambe et al [19], Evangelista [11], [12] Shelby et al [35], and Qiu et al [28]. The first of these reports a class of pitch detection algorithms which are event-based. The authors' approach attempts to determine the GCI (Glottal Closure Instant) as a cue using the multiresolutional properties of wavelets and then estimates the pitch for each sample within a particular segment. The advantages of this approach are that the method does not estimate the average pitch over the period; neither does it assume quasi-stationarity (i.e. short-term

events are better characterised using wavelets). Furthermore, the pitch detector works equally well for speech with either high or low pitch and is robust to noise compared to autocorrelation or cepstral – based pitch detectors which result in higher signal to noise ratios.

Shelby et al [35] reiterate the need for accurate pitch detection in the recognition of tonal languages. Their method builds on the work of [19] where one of the main conclusions was that the GCI method showed promise for pitch period estimation of voiced segments. Their DWT (Discrete Wavelet Transform) algorithm uses a different energy method to locate the onset of an utterance. A comparison with the autocorrelation method using the same energy method for utterance onset detection is reported concluding that the DWT has comparable if not better performance.

The wavelet transform is incorporated into a pitch detection system in [28] and [45], where a method is developed to estimate indirectly the pitch of a signal by first calculating its instantaneous frequency. For successful estimation of the IF, harmonics of the speech signal have to be largely attenuated. Instead of using a conventional time-varying filter for which *a priori* knowledge is required, a bank of bandpass filters of compact time domain support (wavelets) are incorporated into the system, which cover the complete range of pitch frequency.

Evangelista [11] develops the Pitch Synchronous Wavelet Transform (PSWT) as an extension to the Multiplexed Wavelet Transform (MWT) [12], both designed using filter banks. The difference is that the PSWT takes into account variable pitch period. The PSWT reduces to the MWT when the pitch sequence is stationary. Suggested applications include uniquely characterising the behaviour of voiced/unvoiced fricatives, occlusive unvoiced consonants, and vowels via the variation in pitch and time varying structure of the signals.

The MWT is also suitable for representation of what the author terms *pseudo-periodic* signals, those which have pitch. These are a special class of signal which are oscillatory in nature yet have period to period variation i.e. exhibit non-stationarity.

Applications suggested by the author include pitch tracking of speech signals, however the pitch itself must be approximated first using conventional methods.

What the above wavelet representations collectively provide is a multiresolutional analysis of fluctuation from periodic behaviour of the analysed signal. Intuitively this is useful in speech analysis since much of the information is contained in these variations. These representations need further work to render their true potential in speech.

2.3.3 Wavelets in Speaker Identification and Authentication

In order that humans may recognise individual speakers, they must first become familiar with the characteristics of a particular voice. The same is true for machines - the process of getting to know a particular speaker is referred to as training and consists of collecting data from utterances of people to be identified. Acoustic characteristics tend to vary between speakers - the main problem for standard speaker independent recognition systems. Moreover, automatic speaker identifiers try to exploit this variability to characterise individuals. The identification or classification itself (the testing phase), consists of comparing an unidentified utterance to the classifier and making the identification. Also, the manner in which speech features are presented to the classifier i.e. how parametrization is achieved, can significantly affect performance. Commonly used classification techniques for speaker identification are typical of those found in speech recognition systems. They include: Dynamic Time Warping, Hidden Markov Modelling, Vector Quantisation, Auto Regression and Neural Networks. See for more detail O'Shaughnessy [26], Deller [9], or Soong et al [37].

2.4 The Ridge-Skeleton Algorithm

Estimates of frequency and amplitude laws of non-stationary signals are useful for characterising important attributes of speech signals such as formants; resonant frequencies of the vocal tract for a given sound, usually voiced, and pitch; the fundamental frequency of a sound, again usually voiced. The wavelet transform provides the means of extracting instantaneous information by tracking and chaining together events across scales. This method is known as ridge and skeleton extraction.

Delprat et al [30] implements such a scheme using Continuous Wavelet Transform (CWT) derived representations to extract instantaneous frequencies. The CWT is a two dimensional plot which is an overcomplete (redundant) representation of the original signal.

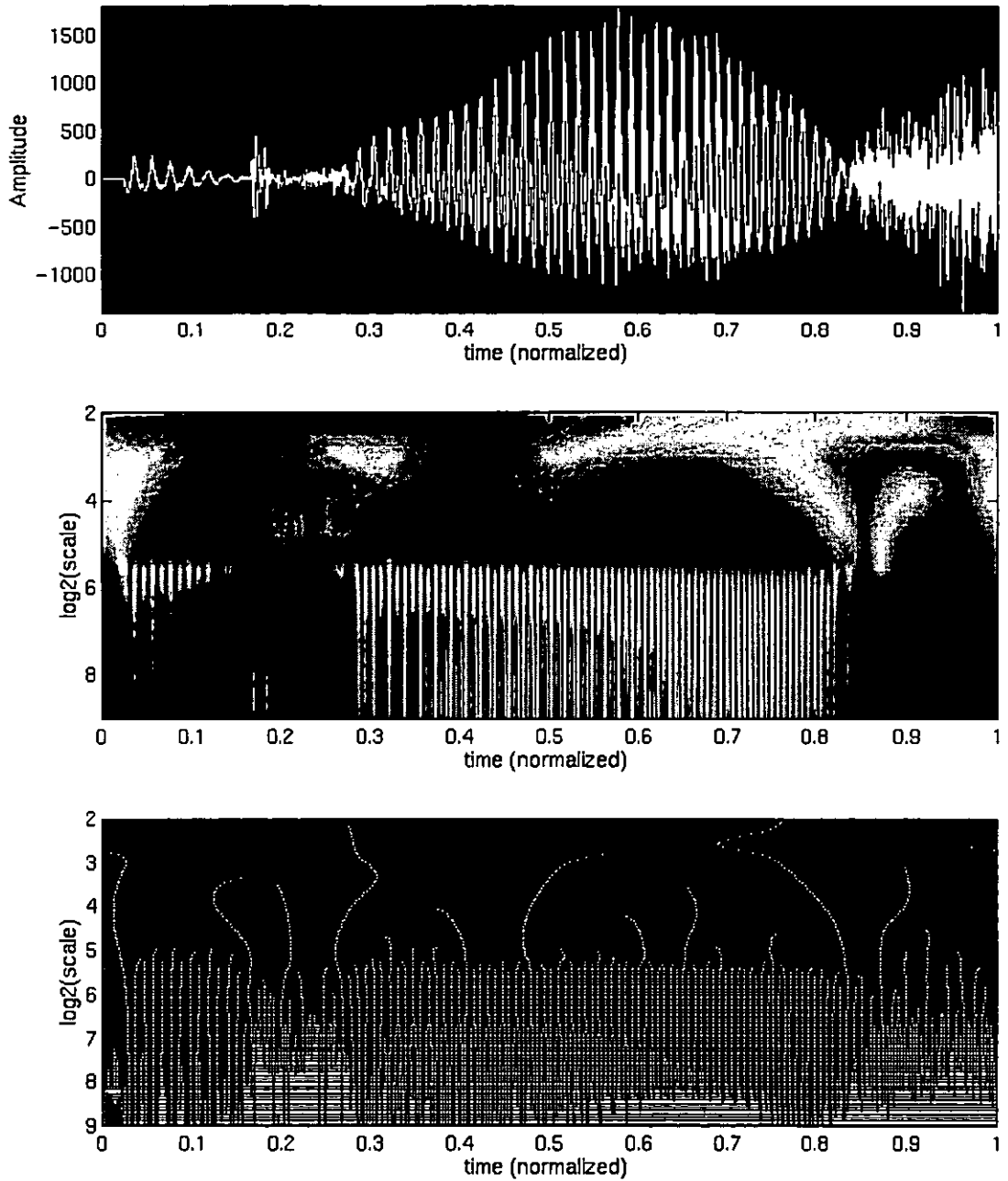


Figure 2.2 Continuous wavelet transform of 'greas' from the word greasy analyzed using a Gaussian wavelet h via the relation $CWT(\tau, a) = \frac{1}{\sqrt{a}} \int f(t)h\left(\frac{t-\tau}{a}\right)dt$ where a is the scale (inversely proportional to frequency) and τ is the translation; Top – Time domain signal; Centre – Continuous wavelet transform and; Bottom – Ridge extraction from the CWT by chaining the maxima modulus of the wavelet amplitudes.

Using the stationary phase method of approximation, the dominant contributions to the wavelet transform are singled out. These are then assigned curves in the time-scale plane (ridges), which are extracted from the modulus maxima of the CWT.

Figure 2.2 shows an example of how ridges may be extracted using the modulus maxima of the continuous wavelet transform for part of the word 'greasy'. Once the

maxima are identified as shown in the figure, they may be chained together to form a skeleton map. This may be pruned to keep only the strongest ridges from which instantaneous frequency or amplitude laws may be extracted.

A possible restriction of this method lies in the fact that the ridge extraction algorithms are not exact for signals such as speech which have a high number of interacting components.

Maes [22] reports success in extracting such features from speech signals using wavelet representations. The method involves transforming the speech signal into a subband decomposition. A 'squeezing' algorithm is then applied which monitors the temporal behaviour of each subband.

Those contributions exhibiting similar temporal behaviour are recombined using a 'fusion' algorithm, effectually building a profile of temporally important speech components. These principal components represent the formants and pitch:- like the ridge-skeleton algorithm, they give information regarding phase and amplitude modulation characteristics of the speech signal. Proposed applications are speaker identification and word spotting. Furthermore, inherent linearity of the wavelet approach ensures robustness to noise although drawbacks due to resolving between closely spaced spectral peaks may occur at higher frequencies because of the worse spectral resolution of the standard wavelet transform.

Considerable reduction in data and the ability to identify a speaker in the presence of noise and/or competing speakers (the cocktail party effect) was reported by Phan et al [27] using wavelet decompositions. A subband coding scheme resulted in a four octave multiresolution decomposition. To prepare the signal for recognition (in this case a template matching scheme), the four octaves were mapped onto a 4 by 64 element matrices, with each row representing a different octave. This format is suitable for the particular recognition scheme used. A 90% reduction in the data necessary for recognition was reported although recognition is considerably degraded at low signal to noise ratios.

Kadambe et al. [21] diversified on previous work with Szu et al.[38], and developed a means of adaptively computing the wavelet transform for feature extraction problems.

The philosophy of the earlier work considered the use of the Adaptive Wavelet Transform (AWT) for classification rather than representation of signals and demonstrated the concept of adaptive wavelets and their applicability to speech.

In the extension of [21], two such applications were considered; the identification of speakers and the classification of unvoiced phonemes. The system first modeled the phonemes using a Daubechies wavelet of order 3 and then attempted to identify a speaker by clustering all the phonemes belonging to the same speaker into one class. A feed-forward neural network architecture was used in the classification stage.

The misclassification rate using this type of classifier was 11%. The adaptive wavelet-based speaker identification system suggested previously in [38] was also considered further with the result that this method was able to identify a given speaker with zero error rate using a very short (one pitch period) segment of speech data. These results, however, were obtained using only three speakers. Further work in this area should consider a larger number of speakers.

2.5 Summary

In this chapter we have reviewed the major areas where wavelets and their variants have found use in speech representation and characterisation. It is generally seen that wavelets are useful tools with many parallels to Fourier methods, however they are able to characterise speech sounds in a richer way. It should be noted that this area is still relatively undeveloped in terms of application to larger speech datasets and in particular, there is little or no development of discriminant wavelet techniques for speech classification. The subsequent chapters will address some of these issues in two main ways

- (i) By providing comprehensive performance statistics of wavelet transforms *cf.* STFT-derived parameters.
- (ii) Investigating discriminant wavelet methods that take advantage of the localised structures within speech.

The following chapter starts by describing wavelet transforms and Fourier transforms and indicates where and why the former may do better. It continues by investigating each techniques performance on several examples of real-world speech data.

2.6 References

- [1] Assaleh, K.T.; Mammone, R.J; Rahim, M.G; Flanagan, J.L; "Speech recognition using the Modulation Model", *Proc. Int. Conf. Acoustics Speech and Signal Processing*, 1993.
- [2] Coifman, R.R. and Wickerhauser M.L. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [3] Coifman, R.R; Saito, N., " Construction of local orthonormal bases for classification and regression.", *Comptes Rendus Acad. Sci. Paris, Serie I* 319 (1994), no.2, 191-196.
- [4] Crochiere, R.E; Weber, S.A; Flanagan, J.L; "Digital Coding of Speech in Subbands," *Bell Systems Technical Journal*, vol.55, pp.1069-1085, Oct.1976.
- [5] D'Alessandro, C and Richard, G, "Random wavelet representation of unvoiced speech," *Proc. IEEE-SP Int.Symp. on Time Frequency and Time Scale Analysis*, pp. 41-44 (Oct.1992).
- [6] D'Alessandro, C., Lienard, L.S., "Decomposition of the speech signal into short time waveforms using spectral segmentation.", *Proc. 1988 IEEE Int. Conf. Acoustics Speech and Signal Processing*, New York, Apr.11-14, 1988, pp.351-354.
- [7] Daubechies, I. "Orthonormal bases of compactly supported wavelets," *Comm in Pure and Applied Math.*, vol.41 No.7, pp.909-996, 1988.
- [8] Davenport, M.R. and Garadudri, H. "A neural net acoustic phonetic feature extractor based on wavelets," *Proceedings of the 1991 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*. Conference Proceedings, Victoria, BC, USA, 09-10 May 1991, (Conf.Code 16372).

- [9] Deller, J.R; Proakis, J.G; Hansen, J.H.L, "Discrete-Time Processing of Speech Signals." New York: Macmillan, 1993.
- [10] Delprat, N; Escudie, B; Guillemain, P; Kronland-Martinet, R; Tchamitchian, P; Torresani, B. "Aysmptotic Wavelet and Gabor Analysis: Extraction of Instantaneous Frequencies," *IEEE Transactions on Information Theory*, vol.38, No.2, March 1992 pp.644-664.
- [11] Evangelista, G. "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Transactions on Signal Processing*, vol. 41, No.12, December 1993.
- [12] Evangelista, G. "Comb and multiplexed wavelet transforms and their application to speech processing," *IEEE Transactions on Signal Processing*, vol. 42, no.2, February 1994.
- [13] Fang, J; McLaughlin, J; Owsley, L; Atlas, L and Sachs, J., "Quadrature detectors for feature extraction in text-independent speaker authentication," *Proc. of the IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 644-647 (Conf. Code 42324).
- [14] Favero, R.E. "Compound Wavelets: Wavelets for speech recognition," *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 600-603 (Conf. Code 42324).
- [15] Gabor, D., "Theory of Communication," *J.IEE*, 93:429 – 457, 1946.
- [16] Gish, H; Schmidt, M., "Text-Independent Speaker Identification.", *IEEE Signal Processing magazine*. pp18-32, Oct. 1994.
- [17] Goldstein, H. "Formant tracking using the wavelet-based DST (Dominant Scale Transform)," *Proceedings of the 1994 IEEE South African Symposium on Communications and Signal Processing - COMSIG-1994*, Stellenbosch, S.Afr, Oct. 4 1994, pp 183-189 (Conf. Code 42995).

- [18] Hoyt, J.D., Weschler, H., "Detection of human speech using hybrid recognition models," *Proc. of the 12th IAPR Int. Conf. on Pattern Recognition*. Part 2 (of 3), Jerusalem, Isr, Oct 9-13 1994 Vol.2 pp.330-333 (Conf.Code 42601).
- [19] Kadambe, S. and Boudreaux-Bartels, G.F. "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [20] Kadambe, S and Srinivasan, P., "Text-independent speaker identification system based on adaptive wavelets," *Proc. of the SPIE - The International Society for Optical Engineering* 1994. Vol. 2242. pp.669-677.
- [21] Kadambe, S and Srinivasan, P. "Applications of adaptive wavelets for speech," *Optical Engineering* 33(7), pp.2204-2211 (July 1994).
- [22] Maes, S. "Nonlinear techniques for parameter extraction from quasi-continuous wavelet transform with application to speech," *Proc. of SPIE - The International Society for Optical Engineering* 1994. Vol.2093 pp. 8-19.
- [23] Malbos; F., Baudry; M., Montresor, S. "Detection of stop consonants with the wavelet transform," *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 612-615 (Conf. Code 42324).
- [24] Malbos; F., Baudry; M., Montresor, S. "Detection of occlusives using the wavelet transform," *Journal de Physique IV* 1994 Vol.4 No.C5 Pt1 pp.493-496.
- [25] Mallat, S. "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Analysis. Machine Intell.*, vol. 31, pp. 674-693, 1989.
- [26] O'Shaughnessy, D. "Speaker Recognition," *IEEE Acoustics Speech and Signal Processing Magazine*, October 1986, pp. 4-17.
- [27] Phan, F; MicheliTzanakou, E; Sideman, S. "Speaker identification with wavelet decomposition and neural networks" *Proc. of the 16th Annual Int. Conf. of the*

- IEEE - Engineering in Medicine and Biology Society*. Part 2 (of 2), Baltimore, MD, USA, Nov 3-6 1994 (Conf.Code 43037), Vol.16 Pt.2 pp.1111-1112.
- [28] Qui, L., Yang, H., Koh, S.N. "Fundamental frequency detector of speech signals based on STFT," *Proc. of the 1994 IEEE Region 10th 9th Annual International Conference (TENCON'94)*. Part 1 (of 2), Singapore, Aug 22-26 1994. Vol. 1. pp.526-529 (Conf. Code 42774).
- [29] Qiu, L.; Yang, H.; Koh, S.N., "Fundamental frequency determination based on instantaneous frequency estimation," *Signal Processing* vol.44 pp.233-241 1995.
- [30] Ramalingam, C.S.; Rao, A.; Kumaresan, R. "Time-frequency analysis using the residual interference signal canceller filter bank," *Proceedings of the IEEE-SP International Symposium on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 500-503 (Conf. Code 42324).
- [31] Reilly, A.P. and Boashash, B. "Comparison of time-frequency signal analysis techniques with application to speech recognition," *Proceedings of the SPIE - The International Society for Optical Engineering*, 1992, Vol.1770, pp.339-350.
- [32] Rioul, O. and Vetterli, M., "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October 1991.
- [33] Saito, N. "Local feature extraction and its application using a library of bases," *Ph.D. Thesis*, Yale University december 1994. Available via ftp from <ftp://pascal.math.yale.edu/pub/wavelets/papers>.
- [34] Saito, N; Coifman, R.R., "Local Discriminant Bases.," *Mathematical Imaging: Wavelet Applications in Signal and Image Processing*. Jul.1994, Proc. SPIE 2303.
- [35] Shelby, G.A.; Cooper, C.M.; Adhami, R.R. "Wavelet-based speech pitch detector for tone languages," *Proc. of the IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 596-599 (Conf. Code 42324).

- [36] Shyuu; J.S., Wang; J.F., Wu, C.H., "Channel-weighting method for speech recognition using wavelet decompositions," *Proc. of the 1994 IEEE Asia-Pacific Conf. on Circuits and Systems*, Taipei, Taiwan, Dec 5-8 1994, pp. 519-523(Conf Code 42903).
- [37] Soong, F., Rosenberg, A., Rabiner, L., Juang, B. "A Vector Quantization Approach to Speaker Recognition." *Proc. Int. Conf. Acoustics Speech and Signal Processing*'85, March 1985, Tampa, pp.387-390.
- [38] Szu, H; Telfer, B; Kadambe, S. "Neural network adaptive wavelets for signal representation and classification," *Optical Engineering*, vol.31 No.9 pp.1907-1916, September 1992.
- [39] Wang; K., Shamma; S.A., Byrne, W.J., "Noise robustness in the auditory representation of speech signals," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, MN, USA, Apr 27-30, 1993, Vol.2, pp.335-338.(Conf. Code 18798).
- [40] Wang, K., and Shamma, S.A., "Modelling the auditory functions in the primary cortex," *Optical Engineering* 1994, Vol.33, No.7, pp.2143-2150.
- [41] Wesfried, E. and Wickerhauser, M.V. "Adapted local trigonometric transforms and speech processing," *Technical Report* (1992) Dauphine University of Paris and Washington University, St. Louis.
- [42] Wickerhauser, M. "Acoustic Signal Compression with Wavelet Packets," In Chui C.K.(ed) *Wavelets: A Tutorial in Theory and applications* (1992).
- [43] Wickerhauser, M. "INRIA Lectures on Wavelet Packet Algorithms." *INRIA*, Roquencourt, France, 1991. Minicourse lecture notes. Available on www at "<http://wuarchive.wustl.edu/doc/techreports/wustl.edu/math/papers/inria300.ps>. Z"
- [44] Yang; X., Wang; K., Shamma, S.A. "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, Vol.38, No.2 pp.824-839, March 1992.

- [45] Yang; H., Qiu; L., Koh, S.N. "Application of instantaneous frequency estimation for fundamental frequency detection," *Proc. of the IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Analysis*, Philadelphia, PA, USA, Oct 25-28 1994, pp. 616-619 (Conf. Code 42324).

Chapter Three

Discrete Wavelet and Fourier Transforms for Phoneme Recognition

3.1 Introduction

This chapter begins by describing fundamental differences between the Fourier and wavelet transforms. In particular, the construction of *frames*, which govern the completeness, stability and redundancy of linear transforms are described for both cases. If one is interested in compression, then it can be shown that the wavelet coefficients are stable in that they converge and decay faster than equivalent Fourier expansions, and are thus a more compact way of representing signals. Non-linear operations carried out in the wavelet domain theoretically outperform Fourier methods in a least squares sense. It is shown that using any one of a number of basis functions shown in Chapter 2, Figure 2.1, when applied to several quite different speech classification tasks, leads to significant improvement in misclassification rates compared to those associated with Fourier derived parameters. This is because in addition to being more efficient approximators, the wavelet transform tiles the

time-frequency plane in a non-uniform way. More specifically, the dyadic wavelet transform has tiling support commensurate with an octave band or constant Q filter bank decomposition (see Figure 3.3). The Short Time Fourier Transform (STFT) on the other hand, uses a window of fixed time-frequency resolution throughout the entire signal range. This indicates the suitability of wavelet methods for speech applications, since a crude model of the acoustical behaviour of the inner ear resolves events containing relatively low frequencies with longer duration time frames in contrast to short transient bursts of high frequency information which are analysed with a higher temporal resolution (see [4] for a general introduction to speech analysis). Wavelets have been used in speech recognition applications to date as described in the previous chapter; however, none to the authors knowledge, have considered the comparative performance of the standard Discrete Wavelet Transform (DWT), with a method as well established and widely used as the STFT. Furthermore, most of the applications of the DWT in speech recognition have been geared toward specific applications such as plosive detection (the presence of other non-transient speech phonemes in the context of other unrelated phonemes). In this study, the standard DWT is applied to various subcategories of speech and other issues are considered, i.e. the choice of basis function. Most applications postulate a suitable basis function based on *a priori* knowledge of the original signal e.g. [7]. While such an approach may be useful for representing a particular signal or class of signals, when one desires features suitable for classification, the best choice of basis function may not be so obvious. In this study, a number of common orthogonal wavelet basis functions are tried on each classification example.

3.2 The Short Time Fourier Transform

One of the most common methods of analysing non-stationary signals [3], is the STFT (also known as the Windowed Fourier Transform). It overcomes the limitations of its long-term counterpart by working on the assumption that if a window, short enough to preserve rapidly changing events in the signal, is chosen then Fourier analysis can be reliably performed. The STFT adds time dependence to the frequency analysis by computing the Fourier coefficients from translated (in time) versions of the window along the signal. The original method, proposed by Gabor [8] used a Gaussian window because it has optimal time-frequency

localisation properties, however it does not provide *compact support* and is known to result in unstable reconstruction. The type of windowing function used really depends on the application, and what is required of it. If one merely wishes to *interpret* reasonably well-localised signal features then reconstruction issues are not a problem and the Gaussian window may well be the best choice. If on the other hand there is some *a priori* knowledge of the signal i.e., if it were known to contain irregular or transient like features of interest, then a window of short-time width (impulse response) would constitute the best choice. Conversely, if the signal in question were well-behaved and smooth, with statistics which varied slowly over time, a longer width of window would suffice. Two problems arise in either circumstance. The first less obvious effect is known as the *Heisenberg inequality principle*. This states that the time-frequency bandwidth product is lower bounded, i.e.

$$\Delta t \cdot \Delta f \geq \frac{1}{4\pi} \quad (3.1)$$

Gabor specified the time-bandwidth product of his Gaussian to be $\Delta t \Delta f = 2\pi$.

The relevance of this becomes clearer if the two-dimensional plot of time versus frequency is considered, i.e. the well known *spectrogram*, with time-frequency tiling as shown in Figure 3.1.

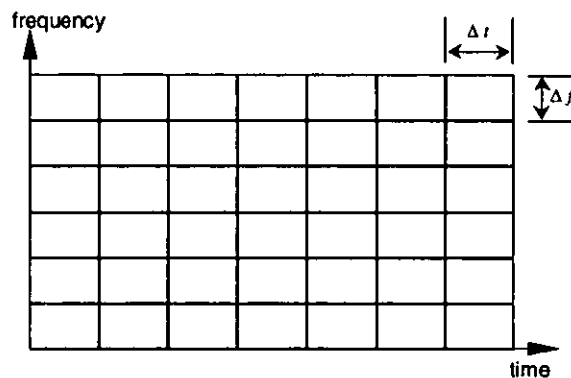


Figure 3.1: Resolution of the time-frequency plane using the Windowed Fourier Transform. Note that this method can only resolve two impulses in time if they are at least Δt apart and two spectral peaks if they are more than Δf apart.

In Figure 3.1, the interpretation of equation (3.1) relates to the area of the time-frequency blocks or *atoms* into which the *t-f* plane is divided. This area is constant regardless of the

rectangular shape of these atoms. In other words, a well localised support in one dimension is at the expense of poorer resolution in the other. Thus good resolution is unattainable in both the dimensions of time and frequency.

The second problem relates more to real-world signals where one might typically encounter signals containing highly non-stationary events, speech is a classic example, where elements such as transients (corresponding to stops and plosives) and well-behaved, almost stationary periodic sections (relating to voicings), silences, etc., are encountered. Ideally, in analysing such a signal, one would wish to choose a particular windowing function for each particular class of sound. The STFT does not allow this: the best window and resolution are chosen at the start of the analysis and from that point onwards, are fixed. In practice, this assumption of *quasi-stationarity* (assuming the signal statistics change slowly, if at all within the window) works well; parameters derived for speech recognition like Fourier derived cepstral coefficients or Fourier derived filter bank coefficients may reflect relatively old technology but they represent a well-understood methodology that, along with Linear Predictive Coding (LPC) techniques have evolved into some of the most popular parameterisation tools in speech recognition systems. However, inherent drawbacks exist. Equation (3.2) shows that the STFT can be considered as the inner product of the original signal $x(t)$ with a modulated window function $g(t)$ centred at $t = \tau$.

$$STFT(\tau, \omega) = \int f(t)g(t - \tau)e^{-i\omega t} dt \quad (3.2)$$

One can easily restrict this expression to a discrete sublattice of the time frequency plane by choosing $\omega_0, \tau_0 > 0$, and defining

$$g_{m,n}^{\omega_0, \tau_0}(t) = e^{-im\omega_0 t} g(t - n\tau_0) \quad (3.3)$$

where m and n are integers. Thus, the windowed Fourier transform coefficients can be computed as the inner product:

$$b_{m,n}(f) = \langle g_{m,n}^{\omega_0, \tau_0}, x \rangle$$

where $\langle \rangle$ denotes the inner product of g and x .

In speech systems $g(t)$ is almost exclusively a *Hamming window* because it has a low spectral leakage due to the high attenuation of its sidelobes although tends to blur in frequency because of the wider bandwidth of the main lobe. It is just a raised cosine function:

$$g_j = 0.54 - 0.46 \cos\left(\frac{2\pi j}{(N-1)}\right) \quad 0 \leq j \leq N \quad (3.4)$$

$$= 0 \quad \text{Otherwise}$$

An alternative interpretation of equation (3.2) can be considered as the projection of $f(t)$ onto a set of building blocks or *basis functions*. The resulting approximation then consists of a weighted series of translated basis functions that in the case of the STFT are just a set of modulated sinusoids possessing constant scale. From the speech analysis viewpoint, it is desirable to have a way of analysing the signals in a *multiresolutional* way.

Another restriction is the limited choice of basis function inherent in the STFT. Fourier methods will most likely provide their best representation on those sections of the speech signal which bear the closest resemblance to the sinusoidal bases ($g(t-\tau)e^{-i\omega\tau}$) themselves. For signals as complex as speech, the STFT is unable to cope with the direct input of raw non-stationary speech and can provide only the best representation when the windowed data is smooth. This issue is discussed in more detail in Section 4.4.

3.3 The Wavelet Transform

There exist many good introductions to the Wavelet Transform, see for example [16] for a good introductory overview of the area or [13] for good tutorial software. In fact, it was not until relatively recently that many existing signal processing techniques such as subband coding, widely used in signal and image compression, pyramid algorithms in image processing, multiresolution in computer vision, and multi-grid methods in numerical

analysis were recognised as being unified within the single common framework of wavelet theory.

The following discussions will focus on the relation of wavelets with the STFT since the Continuous Wavelet Transform (CWT), in particular, is conceptually rather similar.

The CWT can be defined as follows:

$$CWT(\tau, a) = \frac{1}{\sqrt{a}} \int f(t) h\left(\frac{t-\tau}{a}\right) dt \quad (3.5)$$

Where $h\left(\frac{t-\tau}{a}\right)$ is a series of *mother wavelets* or *analysing functions*. In fact, the expression

$\frac{1}{\sqrt{a}} h\left(\frac{t-\tau}{a}\right)$ in equation (3.5) forms a family of shifted, scaled versions of the mother

wavelet with τ corresponding to the translate and a to the scale. It is worth noting that scale and frequency are interdependent, there being an inverse relationship between them. The concept of scale is worth keeping for the time being as it gives an insight into the important multiresolution properties of the wavelet transform. The two main differences from the

STFT lie firstly in that the function $h\left(\frac{t-\tau}{a}\right)$ is not restricted to being sinusoidal. It emerges

that many other families of wavelet exist, each with a particular set of properties. A second difference can be seen from Figure 3.2, which is the wavelet analogue of Figure 3.1.

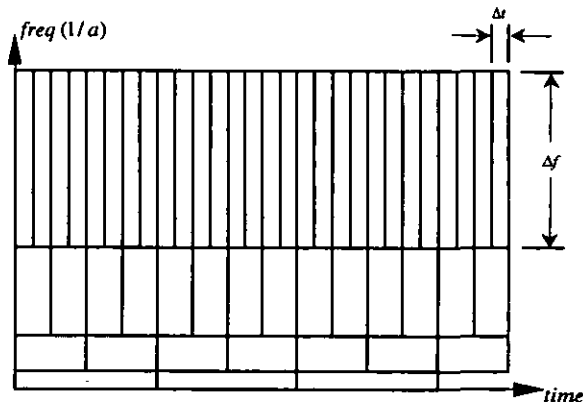


Figure 3.2: Tiling of the Time-Scale plane via the Wavelet Transform. Note the scaling property of the WT that causes the basis function support to grow exponentially with frequency.

The scaling factor a in equation (3.5) given by $a = \frac{\Delta f}{f}$, implies that when f is large, scale is small. Furthermore the presence of a in $h\left(\frac{t-\tau}{a}\right)$ means that as the scale increases, the impulse response of h becomes spread out in time, hence taking only global signal features into account. Because the Heisenberg inequality principle still holds, one achieves good time resolution at low scale (high frequency) at the expense of frequency resolution and vice versa.

3.3.1 Some Fundamental Wavelet Definitions

Orthonormality :- Since the CWT given in equation (3.5) is essentially a mapping from one dimension to two, it is inherently overcomplete. Redundancy may be eliminated by sampling a and τ according to the tiling of the time-frequency plane shown in Figure 3.2, i.e. one sets $a = a_0^m$ and $\tau = nb_0 a_0^m$, where n and m are integers. The resulting wavelets then become

$$h_{m,n}(t) = a_0^{-m/2} h(a_0^{-m} t - nb_0) \quad (3.6)$$

with the added constraint that $\int h(t) dt = 0$. It can be shown that the new h values sampled in this way are orthonormal bases of L^2 space.

The new orthonormal sampled wavelet coefficients now become:

$$c_{m,n}(f) = \langle h_{m,n}, x \rangle \quad (3.7)$$

By analogy with equation (3.3) one can compare the similarity of the two types of basis functions. The sinusoidal terms in the Fourier case are modulated by a window $g(t)$ to form the short time Fourier basis, whereas in (3.6), h forms the family of bases. It emerges that the respective bases of the Fourier and wavelet methods are related as being special cases of a common "Lie-group"; *Weyl-Heisenberg* transforms in the first case and *Affine* transforms in the second. However, there do exist some important differences which, as shall be seen,

indicate the suitability of wavelets over Fourier methods for analysis of real world signals, acoustic or other.

Frames [5] :- In the case of both classes of transform, one can define some restrictions that guarantee stability, that is, given the respective transform coefficients, $b_{m,n}$ and $c_{m,n}$, and the knowledge of the respective basis functions, it is required to define certain conditions within which the expansion will be constrained. If the discretisation parameters a_o , and b_o are known, it is possible to construct the bases $h_{m,n}$ such that they constitute a frame. More accurately, if frame bounds A and B are defined such that for a given basis $h_{m,n}$ then one can write

$$A\|f\|^2 \leq \sum_{m,n} |\langle h_{m,n}, f \rangle|^2 \leq B\|f\|^2 \quad (3.8)$$

where $\|f\|^p = \left(\int_{-\infty}^{+\infty} |f(t)|^p dt \right)^{1/p} < +\infty$, is the norm of f .

If the frame bounds are such that $A = B$, then the sequence $h_{m,n}$ is said to form a *tight frame*, although it does not necessarily form a basis. If $A = B = 1$, then the frame forms an *orthonormal basis*. A frame of this type guarantees a unique representation of any given function in L^2 space. The ratio B/A in general is a measure of transform redundancy – the closer to 1, the faster the convergence. In the case of the STFT, one can achieve tight frames and an orthonormal basis simultaneously; however this gives rise to g 's with poor resolution in frequency or time. If the orthonormality restriction is removed, localisation properties are much improved while tight frames are maintained. This is a result of the *Balian-Low theorem* (see [7] for details). The wavelet case turns out to be quite different, since the choice of h can be such that the set of functions shown in (3.6) automatically forms an orthonormal basis of $L^2(\mathfrak{R})$ if a_o is chosen to have the particular value of 2.

Multiresolution [16], [17] :- An important characteristic of the wavelet transform is that it analyses the signal in question with *constant relative bandwidth*, i.e. the windows vary logarithmically in frequency over the time-frequency plane (Figure 3.2). This property

provides a multiresolutional analysis (MRA) of the signal over $L^2(\mathfrak{R})$ or looked at another way, decomposes $L^2(\mathfrak{R})$ space into a chain of nested subspaces

Resolution increasing \rightarrow

$$\dots \subset V_2 \subset V_1 \subset V_0 \subset V_{-1} \subset V_{-2} \subset \dots \quad (\text{containment}) \quad (3.9)$$

$$\bigcap_{j \in \mathbb{Z}} V_j = \{0\} \quad (\text{guarantees uniqueness}) \quad (3.10)$$

$$\bigcup_{j \in \mathbb{Z}} V_j = L^2(\mathfrak{R}) \quad (\text{guarantees completeness}) \quad (3.11)$$

Equation (3.10) can be interpreted as the intersection of all V_j is empty, equation (3.11) as the union of all V_j is dense i.e. a nested subspace.

Also $V_{-m} \rightarrow L^2(\mathfrak{R})$ as $m \rightarrow \infty$ (approximation approaches original signal as resolution increases) and has the property

$$f(x) \in V_j \Leftrightarrow f(2x) \in V_{j-1}, j \in \mathbb{Z} \quad (\text{scaling property}) \quad (3.12)$$

Given such a set of nested subspaces V_j , there exists another set of subspaces (or detail spaces) W_j which are the orthogonal complements of V_j in V_{j-1} . Thus V_{j-1} can be written:

$$V_{j-1} = V_j \oplus W_j \quad (3.13)$$

Hence W_j can be seen as containing the detail necessary to go from one resolution to the next. As a consequence, $L^2(\mathfrak{R})$ can be spanned by the direct sum of all added details:

$$L^2(\mathfrak{R}) = \bigoplus_{j \in \mathbb{Z}} W_j \quad (3.14)$$

The basic principle of MRA states that if there exists a scaling function which satisfies certain requirements, i.e. smoothness, continuity, and orthonormality, such that

$$\phi_{m,n}(k) = 2^{-m/2} \phi(2^{-m}k - n), \quad n=1,2,\dots, m=0,1,\dots \quad (3.15)$$

forms an orthonormal basis for V_j , then W_j , its orthogonal complement, is similarly spanned by the orthonormal basis

$$\psi_{m,n}(k) = 2^{-m/2} \psi(2^{-m}k - n), \quad n=1,2,\dots; \quad m=0,1,\dots \quad (3.16)$$

i.e. the wavelets in equation (3.16) are essentially the same as in equation (3.6).

It is known from the scaling property (3.12) that any two subspaces, say $V_0, W_0 \subset V_{-1}$ can be generated via integer translates of the scaling functions $\phi(k)$ and $\phi(2k)$ respectively. Given the scaling relation (3.12), there exists a sequence $\{g_j\}$ such that

$$\phi(k) = \sqrt{2} \sum_j g_j \phi(2k - m) \quad (3.17)$$

This equation represents the principal association which determines the multiresolution hierarchy.

Also since $W_0 \subset V_{-1}$, the wavelet (3.16) must satisfy a similar equation:

$$\psi(k) = \sqrt{2} \sum_j h_j \phi(2k - m) \quad (3.18)$$

The filter sequences g_j and h_j are related by :

$$h_j = (-1)^j g_{L-j-1}, \quad j=0,\dots, L-1 \quad (3.19)$$

where g_j is known as the *smoothing* or *scaling* filter and h_j the *detail* or *wavelet* filter. Note that in practice g_j and h_j are usually low and high pass filters which satisfy perfect reconstruction properties. Some other design impositions are applied to g_j from which h_j is obtained by the relation (3.19).

In terms of signal analysis, the concept of MRA can be summarised as follows: If one has an approximation to a given signal at a resolution corresponding to V_j then to go up a resolution, add W_j which contains the detail absent in the lower resolution. Due to the

“ladder property” of the V_j , it follows that the collection of $(\psi_{m,n}; m, n \in \mathbb{Z})$ forms an orthonormal basis of wavelets over $L^2(\mathbb{R})$. See e.g. [3] for details.

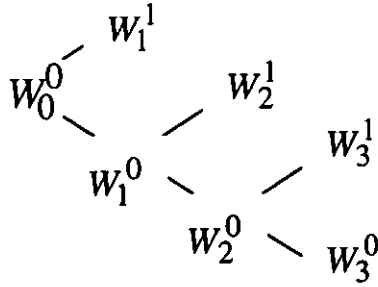


Figure 3.3: A decomposition of W_0^0 ; the parent space is decomposed into mutually orthogonal subspaces using the Discrete Wavelet Transform where the depth J is 3. The final wavelet decomposition is achieved by collecting expansion coefficients present in the terminal nodes.

Translation Invariance :- In pattern recognition applications, it is necessary to have a representation which is translation invariant. This means that the numerical descriptors used for a signal say $f(t)$ should be the same as its shifted equivalent $f(t-u)$ within the given translation u . Transformations like the STFT and the CWT preserve translation invariance but the convolution-subsampling inherent in the orthogonal or discrete wavelet transform destroy it.

For example, the CWT given in (3.5) can be written as:

$$CWT(\tau, a) = \frac{1}{\sqrt{a}} \int f(t) h\left(\frac{\tau-t}{a}\right) dt = f \otimes \tilde{h}(\tau) \quad (3.20)$$

with $\tilde{h}(t) = \frac{1}{\sqrt{a}} h\left(\frac{-t}{a}\right)$. It is therefore independent of translations, e.g. if f is shifted by an amount u , i.e. $f_u(t) = f(t-u)$ then the convolution of (3.20) will not be affected within u

$$CWT_u(\tau, a) = f_u \otimes \tilde{h}(\tau) = CWT(\tau-u, a) \quad (3.21)$$

Forming a wavelet frame as described in Section (3.3.1) samples the wavelet transform uniformly over a dyadic grid. This removes the translation invariance of equation (3.20) as

long as the translation u is not equal to the sampling interval nb_o . The wavelet frames of equation (3.6), i.e.

$$h_{m,n}(t) = a_o^{-m/2} h\left(\frac{t - na_o^m b_o}{a_o^m}\right)$$

sample the continuous wavelet transform according to this interval. With a change of variable $\tau = nb_o a_o^m$ and $a = a_o^m$, $n, m \in \mathbb{Z}$, the DWT can be written

$$DWT(nb_o a_o^m, a_o^m) = \langle f, h_{m,n} \rangle = f \otimes \tilde{h}(nb_o a_o^m) \quad (3.22)$$

$$\text{with } \tilde{h}(nb_o a_o^m) = a_o^{-m/2} h(-nb_o)$$

Once more translating f by u i.e. $f_u = f(t - u)$ yields

$$DWT_u(nb_o a_o^m, a_o^m) = \langle x_u, h_{m,n} \rangle = x \otimes \tilde{h}(nb_o a_o^m - u)$$

The interval at which the wavelet is sampled, $a_o^m b_o$ if large compared with the signal frequency, may result in the coefficients $\langle x, h_{m,n} \rangle$ and $\langle x_u, h_{m,n} \rangle$ being significantly different. In the case of the orthogonal DWT, $a_o^m b_o$ is at a maximum. The translation invariance of the transform is similarly increased.

A number of methods exist ([4], [9] for example) to overcome this problem. Chapter 4 uses a variant of the 'spin-cycle' procedure suggested in [4] to try and remove sensitivity due to translations in the final classification stage. Essentially, the method performs a circulant shift on each of the x_i in the training and testing sets by $-\tau, -\tau + 1, \dots, -1, 1, \dots, \tau$ with $\tau \in \mathbb{N}$, $\tau < n$ where n is the dimension of x . This inevitably leads to an increase in the number of training and testing samples of 2τ extra per signal. The classifier, which uses Linear Discriminant Analysis (LDA) in this case is trained and tested as usual and for each signal plus its shifted versions, a class assignment is thus obtained by taking the majority vote. An implementation of the translation invariant DWT was tried for the experiments discussed in this chapter, the improvement of which was found to be negligible for this particular application. This is

because the sampling interval $2_o^m b_o$ was relatively small compared to the signal frequency (16 Khz bandlimited speech).

The Zoom Property of Wavelets [10],[15] :- A characteristic unique to the wavelet transform is the ability to characterise *local* signal regularity by the decay of the wavelet coefficient amplitude across scales. The Fourier transform, on the other hand can only indicate a function's *uniform* regularity or smoothness. In other words, at high frequencies, regularity cannot be measured at a particular location. *Lipschitz exponents* [10] are the tool used to provide a measure of uniform regularity but in addition, can characterise the behaviour of the signal regularity at a given point (known as pointwise Lipschitz regularity).

If one assumes that $f(t)$ is m times continuously differentiable over an interval $[v-h, v+h]$, then f can be approximated by the Taylor polynomial

$$p_v(t) = \sum_{k=0}^{m-1} \frac{f^{(k)}(v)}{k!} (t-v)^k \quad (3.23)$$

The resulting approximation error for $(v-h) \leq t \leq (v+h)$ is

$$|\varepsilon_v(t)| = |f(t) - p_v(t)| \leq \frac{|t-v|^m}{m!} \quad (3.24)$$

Definition (Lipschitz) [17].

- A function f is pointwise Lipschitz $\alpha \geq 0$ at v if for a constant $C > 0$ we have

$$|f(t) - p_v(t)| \leq K |t-v|^\alpha \quad (3.25)$$

- A function is uniformly Lipschitz α over $[a, b]$ if it satisfies (3.25) for all $v \in [a, b]$ with a constant C independent of v .

The Lipschitz regularity has to do with the local differentiability of a function f in a neighbourhood. If f has a singularity at the point v i.e. it is no longer differentiable then the Lipschitz exponent α characterises this behaviour.

The Lipschitz condition for the uniformly regular Fourier case [17] is as follows:

First recall that the Fourier transform pair can be written as

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(t)e^{-j\omega t} dt \quad (3.26)$$

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \hat{f}(\omega)e^{j\omega t} d\omega \quad (3.27)$$

$f(t)$ is bounded and uniformly Lipschitz α over \mathfrak{R} if:

$$\int_{-\infty}^{+\infty} |\hat{f}(\omega)| (1 + |\omega|^\alpha) d\omega < +\infty \quad (3.28)$$

If ($0 \leq \alpha < 1$), the Taylor expansion of (3.23) simply becomes the first term i.e. $p_\nu(t) = f(\nu)$. In this setting, with the Lipschitz regularity less than 1, the function is actually discontinuous. If (3.28) and (3.25) are combined for this choice of α , one ends up with

$$\frac{|f(t) - f(\nu)|}{|t - \nu|^\alpha} \leq \frac{1}{2\pi} \int_{-\infty}^{+\infty} 2|\hat{f}(\omega)||\omega|^\alpha d\omega = C \quad (3.29)$$

Since (3.28) is satisfied, then $C < \infty$ for all choices of t and ν , implying that f is uniformly regular when seen through a Fourier basis.

Thus the Fourier transform is unable to give information about local or irregular features from the rate of decay of its coefficients, typically faster than $\frac{1}{\omega^\alpha}$.

Wavelets on the other hand can provide this information in a multiresolutional way by indicating the rate of decay across scales. This decay can be directly related to the nature of regularity across and within a signal. Indeed it has been shown in [15] that the measurement of this decay is equivalent to magnifying the signals properties in the neighbourhood of

discontinuities as the scale decreases (frequency increases). The first point to note is that if a high measure of locality using Lipschitz α is required, one needs not just a narrow time window but also a wavelet with a large number of vanishing moments. These vanishing moments are an indicator of the degree to which a particular wavelet is able to approximate the differentiability of f . To see this, the decay of the wavelet coefficients in (3.2) as the scale a decreases has to be considered. For the uniform case, one can measure α directly from the wavelet coefficients via the following upper bound if there exists $S > 0$

$$|DWT(\tau, a)| \leq S a^{\alpha + \frac{1}{2}} \quad (3.30)$$

with $\alpha \leq n$ (the number of vanishing moments).

This proves that $DWT(\tau, a)$ decays faster than $a^{\alpha + \frac{1}{2}}$. If the wavelet has vanishing moments n , the wavelet transform coefficients give no information about the Lipschitz regularity when $\alpha > n$.

To investigate the local regularity in the neighbourhood of a point v using the wavelet transform, it is best to consider a cone of influence about a singularity v on the time-scale plot where the majority of wavelet coefficients lie. It can be shown [15] that the cone of influence is related to the compact support of the wavelet, for example if the support of a wavelet is $[-D, D]$ then in the time-frequency plane for a set of points (τ, a) , the cone of influence is defined as

$$|\tau - v| \leq D a \quad (3.33)$$

where τ is the translation factor of the wavelet. At large scales, this quantity becomes large and at finer scales the regularity is pointwise and well-defined. This is best seen in the scalewise wavelet transform of Figure 3.4, in particular near the singularities which clearly show the cone of influence bounding the high amplitude coefficients.

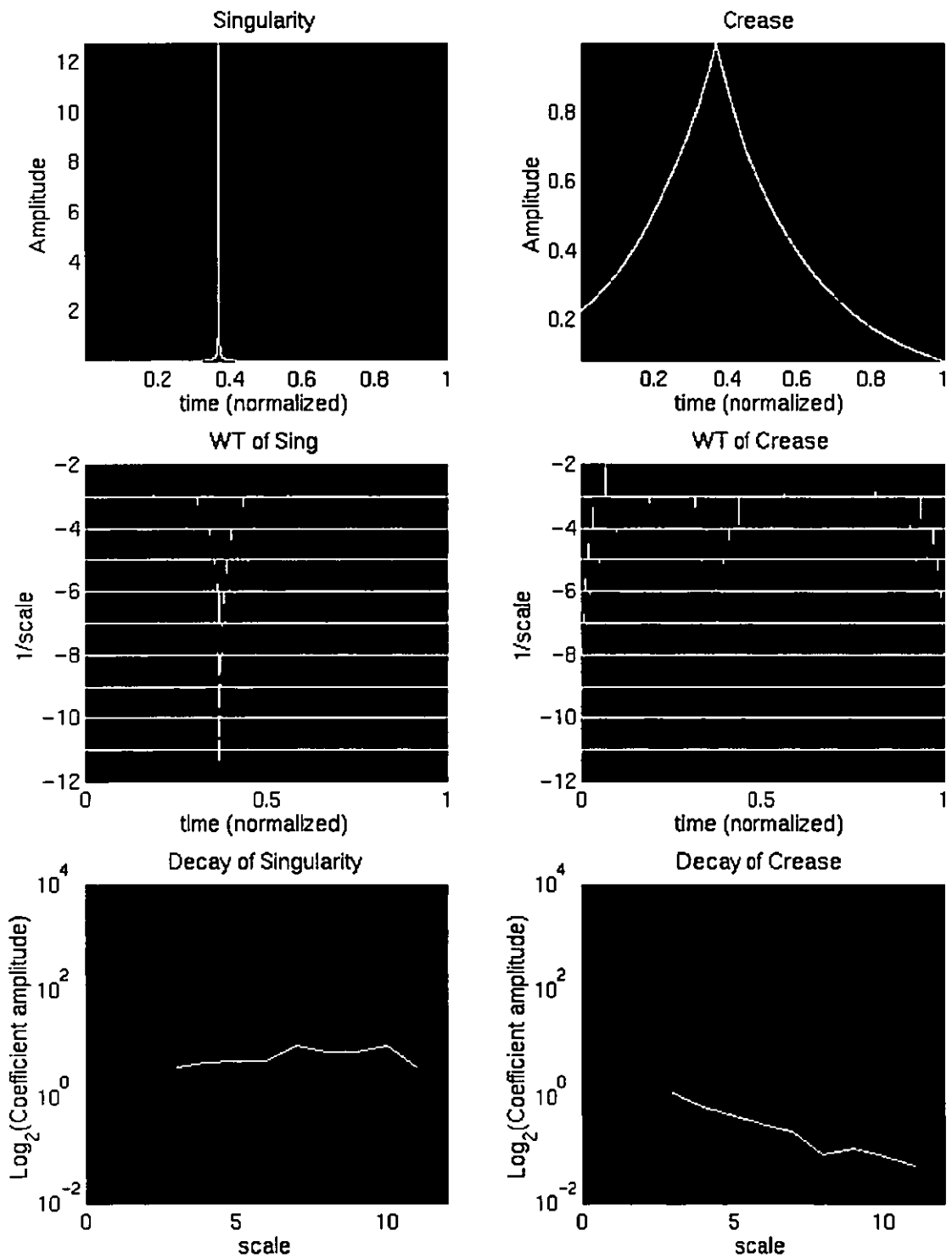


Figure 3.4. Large scales on these plots are those closest to zero; this is where localisation is worst, but increases as the scale becomes finer. Levelwise decay of the wavelet coefficients is given by \log_2 of the wavelet coefficients across scales. The Lipschitz regularity at each singularity can be calculated from the maximum of these slopes.

3.4 Results

In this section, the DWT [equation (3.22)] is applied to several subclasses of speech data which will be described shortly. Basis functions, including those shown in Figure 2.1 were used to transform the respective datasets and the expansion coefficients containing approximately 90% of the signal energy were retained as recognition features. For comparative purposes, an identical operation was carried out on Short-Time Fourier Transform derived features and classification performance similarly obtained. The speech used was 16kHz sampled data extracted from dialect regions one and two of the TIMIT database [20] and included all speakers, both male and female. These dialect regions are geographically close, DR1 corresponds to New England and DR2 to the Northern US. The total number of speakers in DR1 was 49 of which 27% were female. In DR2 the total number of speakers was 102 of which 30% were female giving an overall number of speakers equal to 151 including 50 female speakers. The speech utterances of varying length, which numbered 1512 were extracted from the database and recursively searched to find all instances of each phone used in all its possible contexts. Since the speech was originally 16kHz bandlimited it was segmented into 32ms sections corresponding to 512 samples per signal.

In these experiments, a number of speech subcategories were chosen with which to evaluate DWT performance. The speech data chosen covers examples from the following speech subcategories:- vowels, semivowels, fricatives (both voiced and unvoiced), nasal, and plosive stops. Much of the work carried out to date on speech recognition using wavelets (see for example [1], [11], [12], [13]) have used speech data of between three and five classes, but classes were chosen from across possible categories instead of intra category. For example, the work described in [13] uses adaptive wavelets to classify an unvoiced fricative sound but only within the context of a voiced fricative and a plosive class. This work provides an evaluation of the DWT on the listed classification problems and compares its feature extraction ability with that of the STFT. No pre-processing was done on the speech samples prior to transformation and classification except via the well-known periodogram, commonly used in speech discrimination tasks [18].

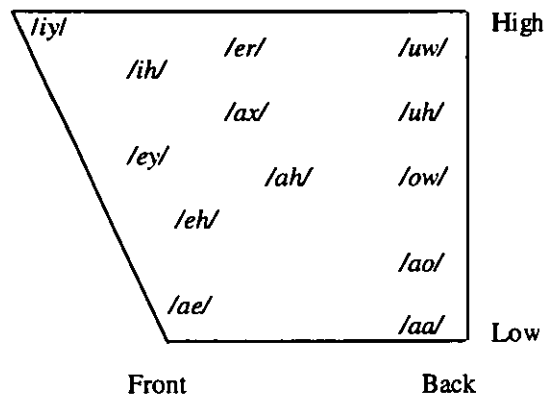


Figure 3.5: Diagram showing distances between several vowel sounds according to the position of the tongue bulk.

Example 3.1 Vowel sounds /iy/, /aa/, and /ax/

The first dataset contained the three vowels /iy/, /aa/, /ax/ corresponding (according to Figure 3.5) to front-, back-, and mid-voiced sounds. If the wavelet transform can efficiently preserve the discriminatory information of these vowels, classification performance should be quite high. Table 3.1 shows the misclassification rates using Linear Discriminant Analysis (LDA) as a classifier when trained using the best 64 expansion coefficients of the transform (section 5.3.2 provides a description of LDA as used throughout this thesis). Clearly the DWT features consistently out-perform STFT-derived parameters regardless of the type of basis function used with the best performance for this problem gained via the 3rd order Deslauriers-Dubuc interpolating wavelet. These wavelets are very symmetric making them suitable for applications like zero-crossing detection and are potentially shift invariant if the transform is redundant (see Donohos paper in [6] for further details).

Since these phonemes are well separated in terms of the vowel triangle, this improvement would indicate that the transform has preserved the formant information. The prediction performance confusion matrix for this example shows that /iy/ and /aa/ have higher confusion with each other than with /ax/ while of the two, /ax/ has higher confusion with /aa/. When compared with the STFT confusion matrix, one can see a similar trend on a greater scale. This could be due to difficulty the STFT has in resolving between the spectral peaks of F1 and F2 which for the phoneme /aa/ are typically close together compared to /iy/. Since the speech samples were acquired across a large range of speakers (both male and

female), it is also probable that overlap has occurred due to the range of pitch and intonation likely to exist between speakers, particularly those of different gender.

Example 3.2 Unvoiced plosives /p/,/t/,/k/

Plosives represent transient sounds generated by the build up of compressed air by causing a total constriction at various points in the vocal tract: (i) At the lips-*bilabial* as in /p/ and /b/, (ii) At the tip of the tongue *alveolar* as in /t/ or /d/, or (iii) towards the back of the vocal tract *velar*, e.g. /k/ or /g/. In the case of the unvoiced stops, typical examples of which are shown in Figure 3.7, there is a period of noisy friction after release due to the sudden turbulence of the released air. This period is followed by a more steady flow of air called aspiration which is still noisy but where some modulation is evident. These two events occur just before vocal fold movement is required for articulation of the next voiced sound and after a *stop gap* which occurs prior to release and friction. All of these features are evident in the time waveforms of Figure 3.7 and their corresponding multiresolutional representations derived via relation (3.9). Because of this property of wavelets, one would expect better performance in this instance compared with standard Fourier methods due to their better characterisation of transients, indeed a significant amount of work has been focused on enhanced representation of signals containing different kinds of singularities (see [14] for an example). In the present experiment, the worst overall misclassification rate was attributed to the top 64 STFT-derived expansion coefficients, while the best, showing almost a 19% reduction in this figure was the DWT having a triangular interpolating mother wavelet. The DWT classification shows confusion particularly in distinguishing /p/ and /k/; however, overall improvement is marked when compared with the STFT confusion matrix (Table 3.2).

Example 3.3 Nasal Stops /m/,/n/,/ng/

These consonants are voiced and similar to vowels except the oral cavity is closed or partially closed while the nasal cavity is open. They are normally of weaker energy than vowels due to the inability of the nasal cavity to radiate sound. They can also, like the plosive stops be characterised by whether they are labial as in /m/, closure made at the lips, alveolar with the tip of the tongue resting just behind the top gum as in /n/, or velar as in /ng/ with the middle or back of the tongue resting on the soft palate.

The confusion matrices of Table 3.3 show an overall prediction misclassification rate of nearly 55% in the case of the STFT falling to around 46% in the case of the Deslauriers-Dubuc wavelet transform. This overall poor performance could probably be attributed to analogy with the human auditory system which itself has difficulty in discriminating these nasal sounds due to the presence of spectral nulls caused by antiresonances in the overall vocal system. The STFT confusion matrix shows even more misclassified /ng/ phones as /n/ phones. Improvement could probably be gained in all cases by providing an active pre-emphasis processing stage to boost the overall energy of the nasals.

Example 3.4 Unvoiced fricatives /f/, /T/, /s/

The set of unvoiced fricatives are generated by creating a turbulent airflow at some point of constriction in the vocal tract. *Labiodental* as in /f/ causes the sound by creating friction between the top teeth and the lower lip. Forcing airflow between the top teeth and the tip of the tongue as in the 'th' sound of thing (/T/) is known as *interdental* and where articulation takes place between the tip of the tongue and the gum is called *alveolar*, an example of which is /s/ as in sing. The main characteristic observable from Figure 3.9 is that the acoustic waveforms are generally noisy and of high frequency, although both /f/ and /T/ have a more burst-like start which is well characterised in the MRA analysis directly beneath each plot. Perhaps it is this attribute that causes more /f/ and /T/ sounds to become confused in the confusion tables of Table 3.4. The simpler Haar wavelet, normally the most suitable for describing discontinuous functions provides best performance here contrary to the work carried out in [13] which suggests that a noisy type of mother wavelet (in fact they use a Daubechies order 3 wavelet) should be used for noisy speech sounds exactly like these. Admittedly the Daubechies wavelet with 7 vanishing moments has comparable performance. Overall misclassification improvement compared with the STFT is around 27%.

Example 3.5 Voiced fricatives /v/, /dh/, /z/

Voiced fricatives contain *mixed excitation* in that they include a voiced component in addition to the original friction. Otherwise the phonemes tested here; /v/, /dh/, /z/ have a similar place of articulation to their unvoiced counterparts /f/, /T/, /s/ considered previously.

As one would expect, this results in an element of periodic modulation in the acoustical waveform, evident in Figure 3.10 particularly for /v/ and /dh/. The phone /z/ is more noiselike in its acoustical content. On observation of the confusion matrices in Table 3.5 one sees higher confusion in both instances (wavelet and Fourier) whilst attempting to differentiate between /v/ and /dh/ compared with /z/ where least errors are made. The best overall performance is gained by using the DWT with a 10 vanishing moment Coiflet wavelet which was around 18% cf. STFT 26%.

Example 3.5 Semivowels /w/, /y/, /l/, /r/

Semivowels are classified into either liquids: /w/ as in work and /l/ as in laugh, or glides: /y/ as in yoke, /r/ as in rum. Glides have a transient formant transition that moves away from some target position. Compared to vowels, which spend much longer at a steady state position, glides spend more time in transit. Liquids are also similar to vowels but have a lower associated energy, Figure 3.11. Table 3.6 shows respective misclassification rates for this group with an improvement from 33.72% error in the case of the STFT to 24.33% in the case of the triangular interpolating wavelet. Of the four classes, /w/ and /l/ exhibit the largest amount of confusion while /y/ and /r/ are relatively well-separated.

Example 3.6 The stressed vowel set /iy/, /ih/, /eh/, /ey/, /ae/, /aa/, /aw/, /ay/, /ah/, /ao/, /oy/, /ow/, /uh/, /uw/, /ux/, /er/.

In the final example, all sixteen of the stressed vowel set were used to assess the robustness of DWT features when applied to a larger range of classes. The results are shown in Table 3.7 and 3.8 where we see a nearly 10% reduction in misclassification compared to the STFT. Note that the results from Table 3.1 were used to decide which wavelet filters were most suitable for vowels to inform the choice of wavelet filter used here. The confusion matrix has been given a different format here; diagonal elements representing testing performance for each vowel while off diagonal elements indicate the Mahalanobis distances (as defined in equation (5.12)) between the wavelet features of the respective classes. The distances can be seen to corroborate somewhat those indicated in Figure 3.5.

3.5 Discussion

The following points arise from the experiments described above:

(i) The wavelet transform is a good estimator of local signal attributes. Fourier methods are powerful tools for global measures but in speech signals, most of the information is carried in irregular signal structure and transitory features. In the wavelet transform, one can characterise local signal regularity by the decay of the wavelet coefficient amplitude across scales as described in Section 3.3.1. This is evidently a useful attribute for the representation of our signals. Section 3.3.1 also relates the vanishing moments of a basis function to the regularity of the signal; overall in the results it is seen that the basis functions with the higher vanishing moments do best in terms of classification and this factor is most likely achieved through more efficient extraction of localised time-frequency features.

(ii) In particular, phonemes with transitory components are analysed well with wavelets of higher vanishing moments. This again has to do with the pointwise regularity of these signals; high amplitude wavelet coefficients existing within the cone of influence of transients provide a better representation when compared with Fourier methods.

It should also be noted that undoubted improvement could be gained by using a more sophisticated e.g. non-linear regime, for classification instead of LDA. However the purpose of this chapter was to provide an indication of the efficacy of the wavelet features for a given application rather than gaining the highest performance per se. The only reason that using a different method of classification might be of interest would be to assess whether the resultant set of features were oblique, in other words would different class information would come to light if the same features were viewed through a different classifier? This is an issue considered further in Chapter 5.

The results given here also motivate 'perceptual' wavelet schemes. We have tried a comparison between the classification results here and the mel-frequency derived cepstral coefficients widely used as parameterisation techniques in speech recognition systems and found them still to be inferior to that gained via the DWT.

3.6 Summary

In this chapter the suitability of wavelet-derived features for phoneme classification have been considered and an overall improvement in performance compared with the short-time Fourier transform (STFT) was gained. No *a priori* assumptions regarding the speech signals were made, the expansion coefficients for each transform were ranked according to their energy and about the top 10% selected as suitable features for recognition. Since a large selection of speakers were chosen (both male and female over more than one dialect region) it was possible to assess the relative robustness of the DWT across a range of common phoneme classification problems. It maintained a significant improvement over the STFT in all cases. This motivates further investigation into the method and its generalisations – this will be the topic of Chapter 4 in which adaptive multiresolutional wavelets are implemented and tried out on speech data.

Technique		MisClassification Rate (%)
STFT_64	Tr	16.48
	Pr	22.38
D6_64	Tr	9.40
	Pr	8.81
D7_64	Tr	10.10
	Pr	10.78
D8_64	Tr	10.34
	Pr	9.65
C6_64	Tr	9.16
	Pr	9.65
C8_64	Tr	9.27
	Pr	9.80
C10_64	Tr	9.70
	Pr	9.52
Haar_64	Tr	9.40
	Pr	9.66
S6_64	Tr	9.10
	Pr	9.37
S7_64	Tr	9.39
	Pr	10.35
S8_64	Tr	10.10
	Pr	9.51
DD3_64	Tr	9.10
	Pr	7.55
CDF_3_9	Tr	9.10
	Pr	9.37
Triangle	Tr	9.15
	Pr	7.69

Table 3.1 Showing misclassification rate of standard DWT algorithm on the unvoiced plosives- /iy/, /aa/, /ax/ and associated confusion matrices using the best of several possible basis functions (see Key).

iy	aa	ax
575	341	777

iy	aa	ax
249	128	338

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

		Predicted Class		
		iy	aa	ax
True Class	iy	223	20	6
	aa	6	119	3
	ax	6	13	319

LDA Confusion Matrices for Testing dataset using DWT Coefficients derived via the third order Deslauriers-Dubuc Interpolating Wavelet.

		Predicted Class		
		iy	aa	ax
True Class	iy	177	58	14
	aa	40	70	18
	ax	10	20	308

LDA Confusion Matrix for Testing datasets using the STFT.

NB. For the basis functions abbreviations are:

C = Coiflet;

D = Daubechies;

S = Symmlet;

DD3-Deslauriers-Dubuc Interpolating wavelet order 3

CDF3_9=Cohen-Daubechies-Feauveau

Biorthogonal Symmetric wavelet

Triangle = Triangular Interpolating wavelet.

In all other cases, the number immediately following the acronym is the number of Vanishing moments

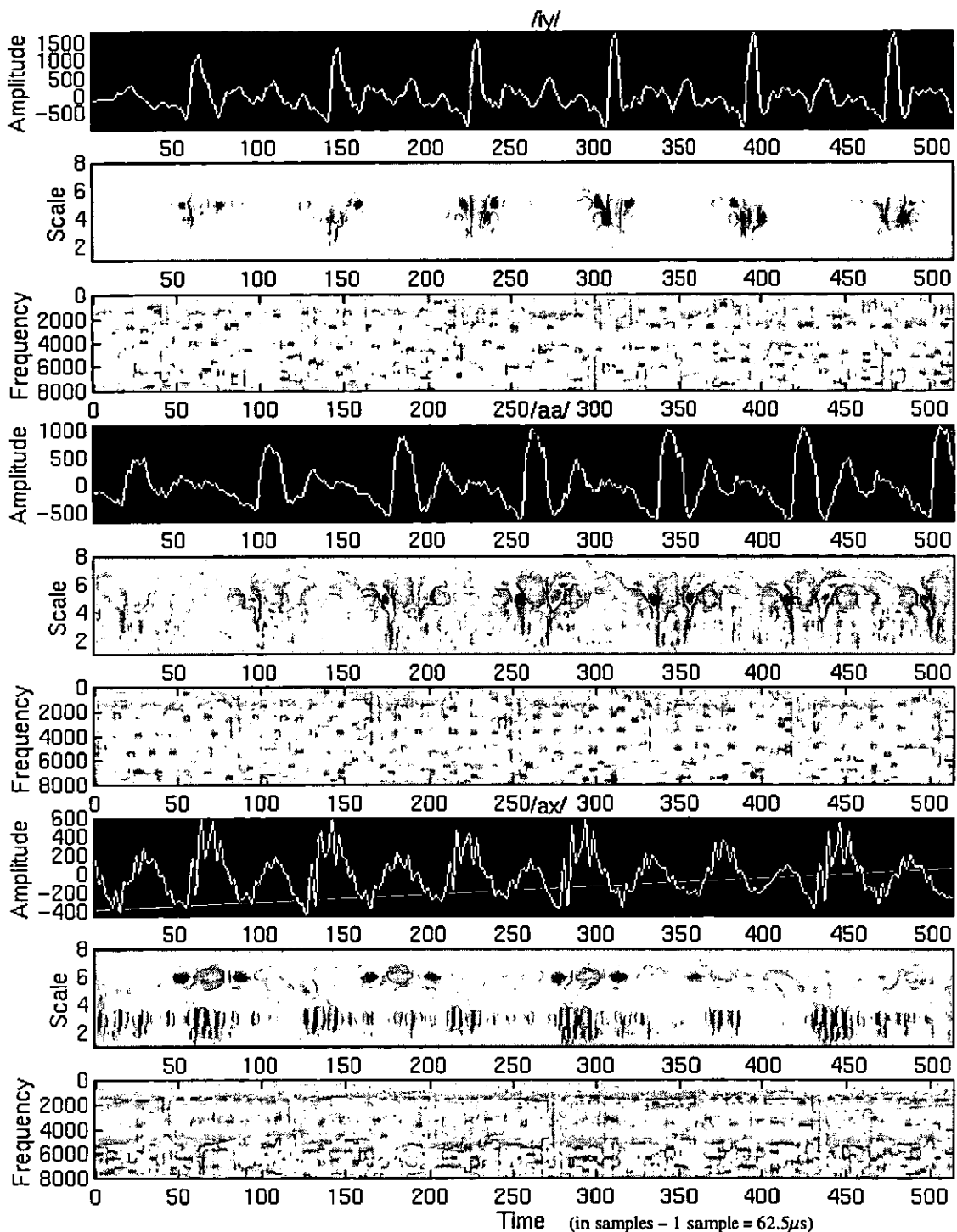


Figure 3.6: Showing typical vowel sounds as used in Example 3.1. The plot directly below each time series shows the multiresolutional analysis of the signal with time (x-axis) versus scale (y-axis). Note that scale is inversely proportional to frequency and so high frequency (transitory) features correspond to low scales while the slowly varying low pass trends are encapsulated in the high scales. The scales are also mutually orthogonal as described in Section 3.3.1. The STFT below the MRA plots is included for comparison. Note the averaging effect of the Heisenberg windows at the higher frequencies (cf. lower scales). Although the STFT provides the ability to track frequencies of interest in time, it is poor at representing transients or showing the behaviour of a feature across frequencies unlike the DWT.

Technique		MisClassification Rate (%)
STFT_64	Tr	39.16
	Pr	46.13
D6_64	Tr	32.25
	Pr	40.65
D7_64	Tr	31.21
	Pr	39.35
D8_64	Tr	30.58
	Pr	42.58
C6_64	Tr	31.41
	Pr	42.10
C8_64	Tr	30.58
	Pr	39.03
C10_64	Tr	30.88
	Pr	42.26
Haar_64	Tr	31.00
	Pr	43.23
S6_64	Tr	31.73
	Pr	40.00
S7_64	Tr	32.36
	Pr	39.68
S8_64	Tr	32.36
	Pr	39.35
DD3_64	Tr	29.53
	Pr	38.71
CDF_3,9	Tr	29.95
	Pr	39.04
Triangle	Tr	29.11
	Pr	38.06

Table 3.2 Showing misclassification rate of standard DWT algorithm on the unvoiced plosives- /p/,/t/,/k/ and associated confusion matrices using the best of several possible basis functions.

p	t	k
224	368	363

p	t	k
60	122	128

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

		Predicted Class		
		p	t	k
True Class	p	39	7	14
	t	20	80	22
	k	36	19	73

LDA Confusion Matrices for Testing dataset using DWT Coefficients derived via the Triangular Interpolating Wavelet.

		Prediction Misclassification		
		p	t	k
True Class	p	33	12	15
	t	14	80	28
	k	30	43	54

LDA Confusion Matrices for Testing datasets using the STFT.

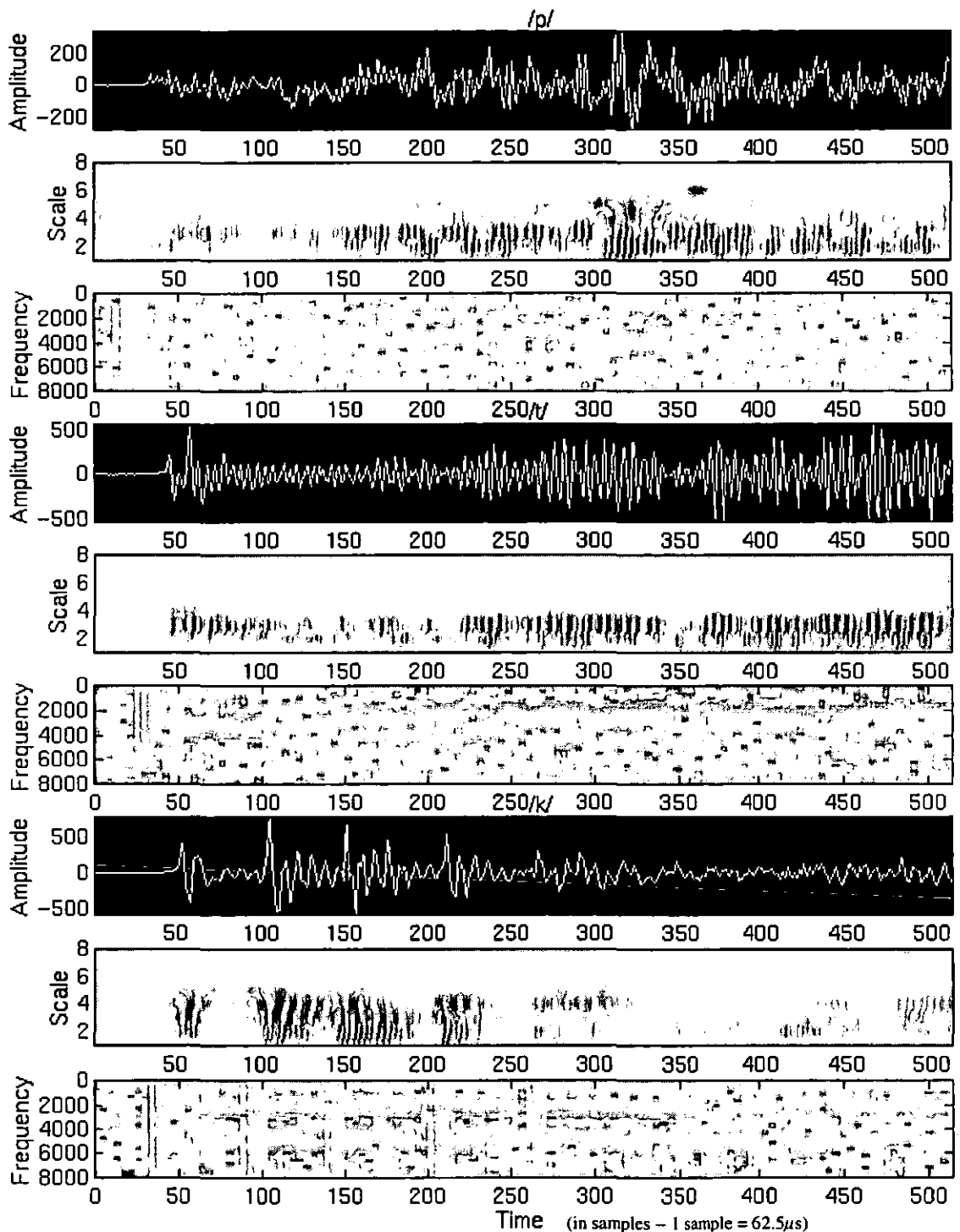


Figure 3.7: Showing the transitory unvoiced plosive sounds classified in Example 3.2. Note the good time localisation of the burst like features where white correspond to high positive amplitudes and black to large negative values. This is in contrast to the STFT plots which shows this method is particularly poor at representing the local effects present in these types of phoneme.

Technique		MisClassification Rate (%)
STFT_64	Tr	52.71
	Pr	54.76
D6_64	Tr	46.05
	Pr	48.58
D7_64	Tr	44.43
	Pr	48.22
D8_64	Tr	41.80
	Pr	46.40
C6_64	Tr	45.10
	Pr	46.99
C8_64	Tr	45.87
	Pr	48.73
C10_64	Tr	45.35
	Pr	48.29
Haar_64	Tr	43.71
	Pr	49.24
S6_64	Tr	46.59
	Pr	49.38
S7_64	Tr	42.57
	Pr	47.50
S8_64	Tr	45.86
	Pr	48.66
DD3_64	Tr	41.98
	Pr	46.11
CDF_3,9	Tr	42.98
	Pr	46.40
Triangle	Tr	41.74
	Pr	46.55

Table 3.3 Showing misclassification rate of standard DWT algorithm on the nasals- /m/,/n/,/ng/ and associated confusion matrices using the best of several possible basis functions.

m	n	ng
1264	2024	421

m	n	ng
501	708	166

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

True Class	Predicted Class		
	m	n	ng
m	314	118	69
n	184	344	180
ng	31	52	83

LDA Confusion Matrices for the Testing dataset using DWT Coefficients derived via a third order Deslaurier-Dubuc Wavelet.

True Class	Prediction Misclassification		
	m	n	ng
m	278	133	90
n	135	273	300
ng	22	73	71

LDA Confusion Matrices for the Testing dataset using the STFT.

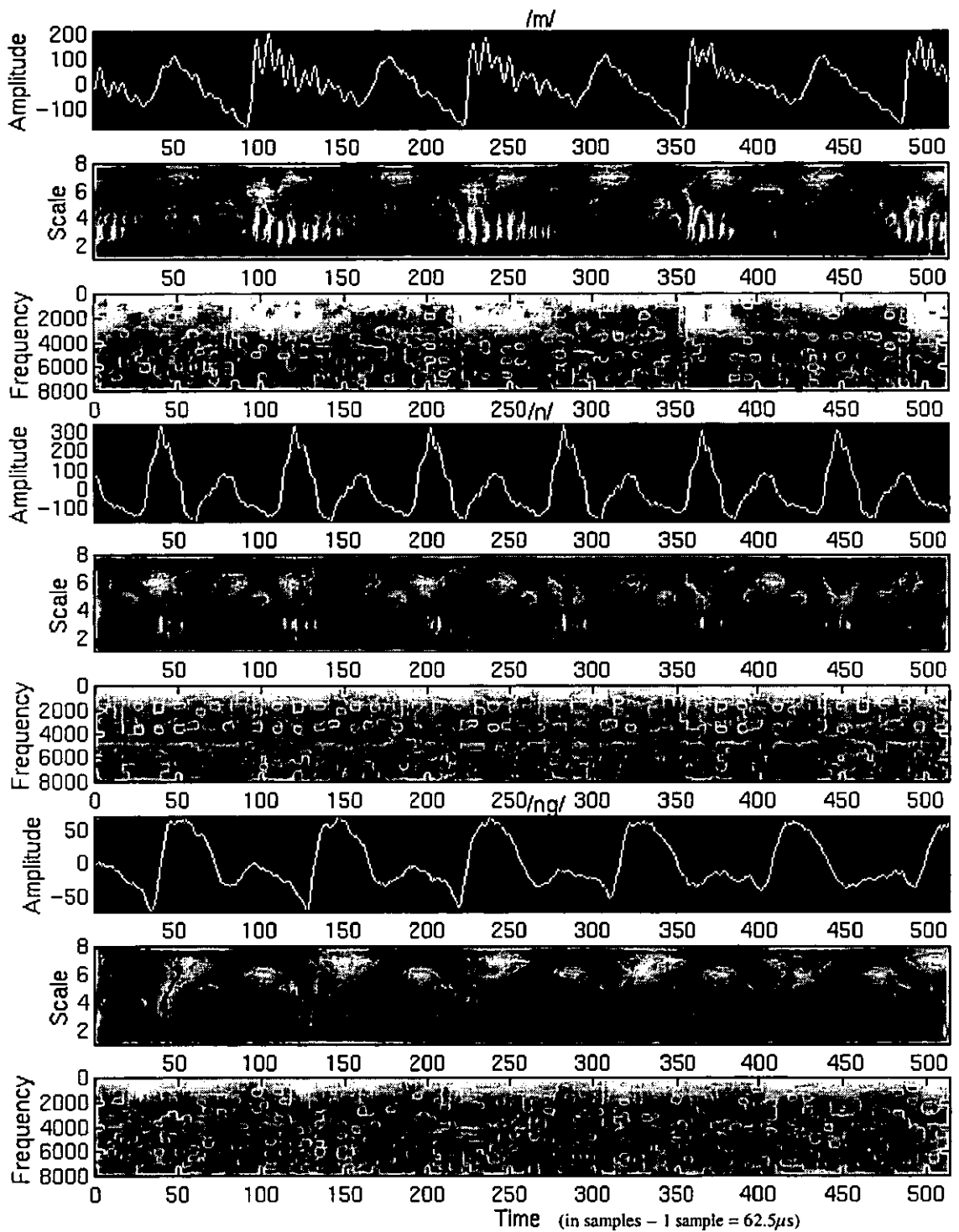


Figure 3.8: Time, MRA and STFT plots of the nasals sounds described in Example 3.3. Again the STFT captures global events but is unable to simultaneously describe the irregular, more dynamic structures compared with the DWT.

Technique		MisClassification Rate (%)
STFT_64	Tr	23.53
	Pr	23.54
D6_64	Tr	17.69
	Pr	18.44
D7_64	Tr	17.82
	Pr	17.43
D8_64	Tr	17.38
	Pr	17.91
C6_64	Tr	17.76
	Pr	18.31
C8_64	Tr	17.54
	Pr	17.34
C10_64	Tr	17.85
	Pr	17.95
Haar_64	Tr	17.68
	Pr	17.25
S6_64	Tr	17.77
	Pr	18.39
S7_64	Tr	17.74
	Pr	17.64
S8_64	Tr	17.72
	Pr	17.6
DD3_64	Tr	17.48
	Pr	18.04
CDF_3,9	Tr	17.77
	Pr	18.39
Triangle	Tr	17.33
	Pr	17.87

Table 3.4 Showing misclassification rate of standard DWT algorithm on the unvoiced fricatives- /f/,/T/,/s/ and associated confusion matrices using the best of several possible basis functions.

f	T	s
1416	416	4228
f	T	s
583	107	1583

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

		Prediction Misclassification		
		f	T	s
True Class	f	415	137	31
	T	31	68	8
	s	89	96	1398

LDA Confusion Matrices for Testing dataset using DWT Coefficients derived via the basic Haar Wavelet.

		Prediction Misclassification		
		f	T	s
True Class	f	327	226	30
	T	52	51	4
	s	138	85	1360

LDA Confusion Matrices for Testing dataset using the STFT.

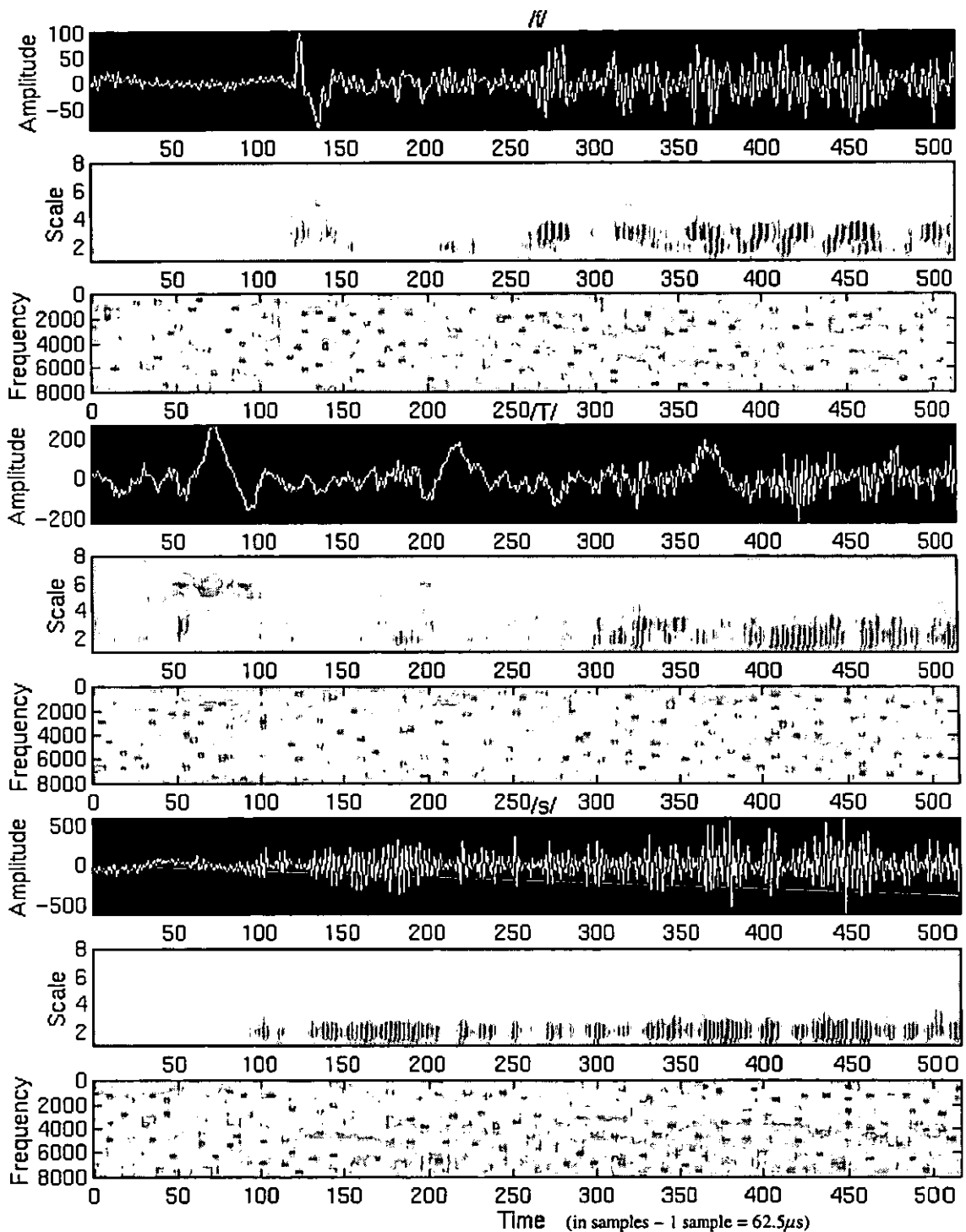


Figure 3.9 Showing time, MRA and STFT plots of three typical unvoiced fricatives as used in Example 3.4. Note that they are nearly all high frequency and noise-like in structure. This is exhibited in the lower scales which once more is difficult to see in the time-frequency STFT diagrams.

Technique		MisClassification Rate (%)
STFT_64	Tr	23.84
	Pr	26.18
D6_64	Tr	17.67
	Pr	18.81
D7_64	Tr	17.09
	Pr	18.43
D8_64	Tr	16.75
	Pr	18.62
C6_64	Tr	17.09
	Pr	19.10
C8_64	Tr	16.92
	Pr	17.86
C10_64	Tr	17.43
	Pr	17.77
Haar_64	Tr	17.53
	Pr	18.24
S6_64	Tr	17.29
	Pr	18.81
S7_64	Tr	16.44
	Pr	18.72
S8_64	Tr	16.41
	Pr	18.91
DD3_64	Tr	16.75
	Pr	18.81
CDF_3,9	Tr	16.68
	Pr	19.00
Triangle	Tr	16.23
	Pr	18.62

Table 3.5 Showing misclassification rate of standard DWT algorithm on the voiced fricatives- /v/,/dh/,/z/ and associated confusion matrices using the best of several possible basis functions.

v	dh	z
705	520	1707
v	dh	z
264	170	624

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

Prediction Misclassification				
True Class		v	dh	z
	v	179	78	7
	dh	51	113	6
	z	18	23	583

LDA Confusion Matrices for Testing dataset using DWT Coefficients derived via the Coiflet filter with 10 vanishing moments (C10).

Prediction Misclassification				
True Class		v	dh	z
	v	132	117	15
	dh	68	92	10
	z	32	35	557

LDA Confusion Matrices for Testing datasets using the STFT.

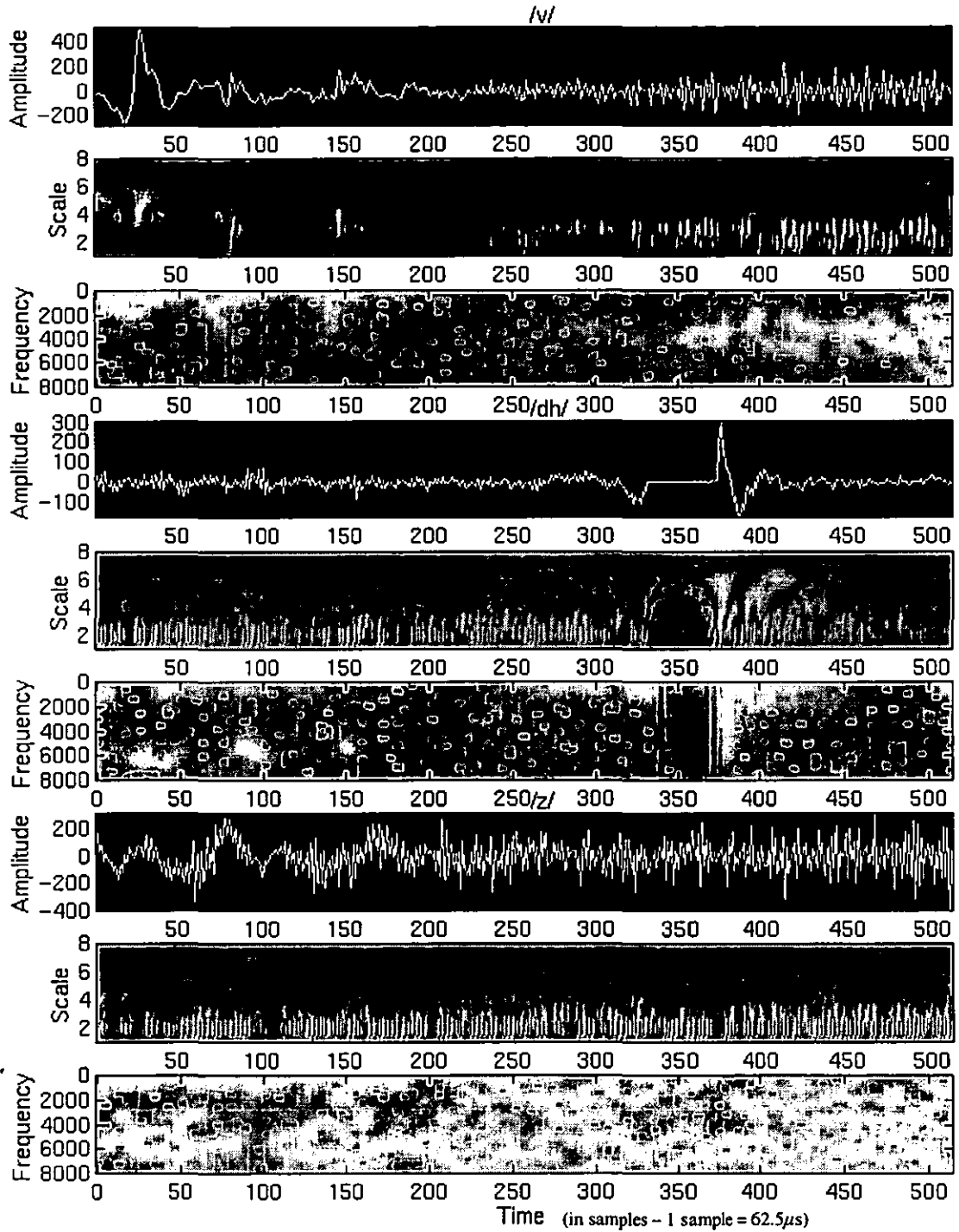


Figure 3.10: Three typical voiced fricatives as used in Example 3.5. The voicing gives a periodic lower frequency element to the signal which is noticeable in the higher scales of the MRA plots. This effect is less visible in the STFT plots.

Technique		MisClassification Rate (%)
STFT_64	Tr	32.82
	Pr	33.73
D6_64	Tr	21.07
	Pr	26.47
D7_64	Tr	21.61
	Pr	26.77
D8_64	Tr	21.36
	Pr	27.54
C6_64	Tr	21.54
	Pr	26.47
C8_64	Tr	21.66
	Pr	26.77
C10_64	Tr	21.28
	Pr	26.47
Haar_64	Tr	21.81
	Pr	26.12
S6_64	Tr	21.50
	Pr	26.90
S7_64	Tr	21.10
	Pr	25.58
S8_64	Tr	21.57
	Pr	27.25
DD3_64	Tr	20.19
	Pr	24.63
CDF_3,9	Tr	21.71
	Pr	26.72
Triangle	Tr	20.36
	Pr	24.33

Table 3.6 Showing misclassification rate of standard DWT algorithm on the semivowels- /w/,/y/,/l/,/r/ and associated confusion matrices using the best of several possible basis functions.

w	y	l	r
801	314	1639	1461

w	y	l	r
317	123	664	577

Frequencies at which phonemes occur in training (top) and testing (bottom) sets.

Prediction Misclassification					
True Class		w	y	l	r
	w	222	17	43	35
	y	7	104	6	6
	l	86	32	476	70
	r	44	20	43	470

LDA Confusion Matrices for Testing dataset using DWT Coefficients derived via the filter with 10 vanishing moments (C10).

Prediction Misclassification					
True Class		w	y	l	r
	w	197	28	56	36
	y	4	108	5	6
	l	130	69	374	91
	r	43	47	52	435

LDA Confusion Matrices for Testing datasets using the STFT.

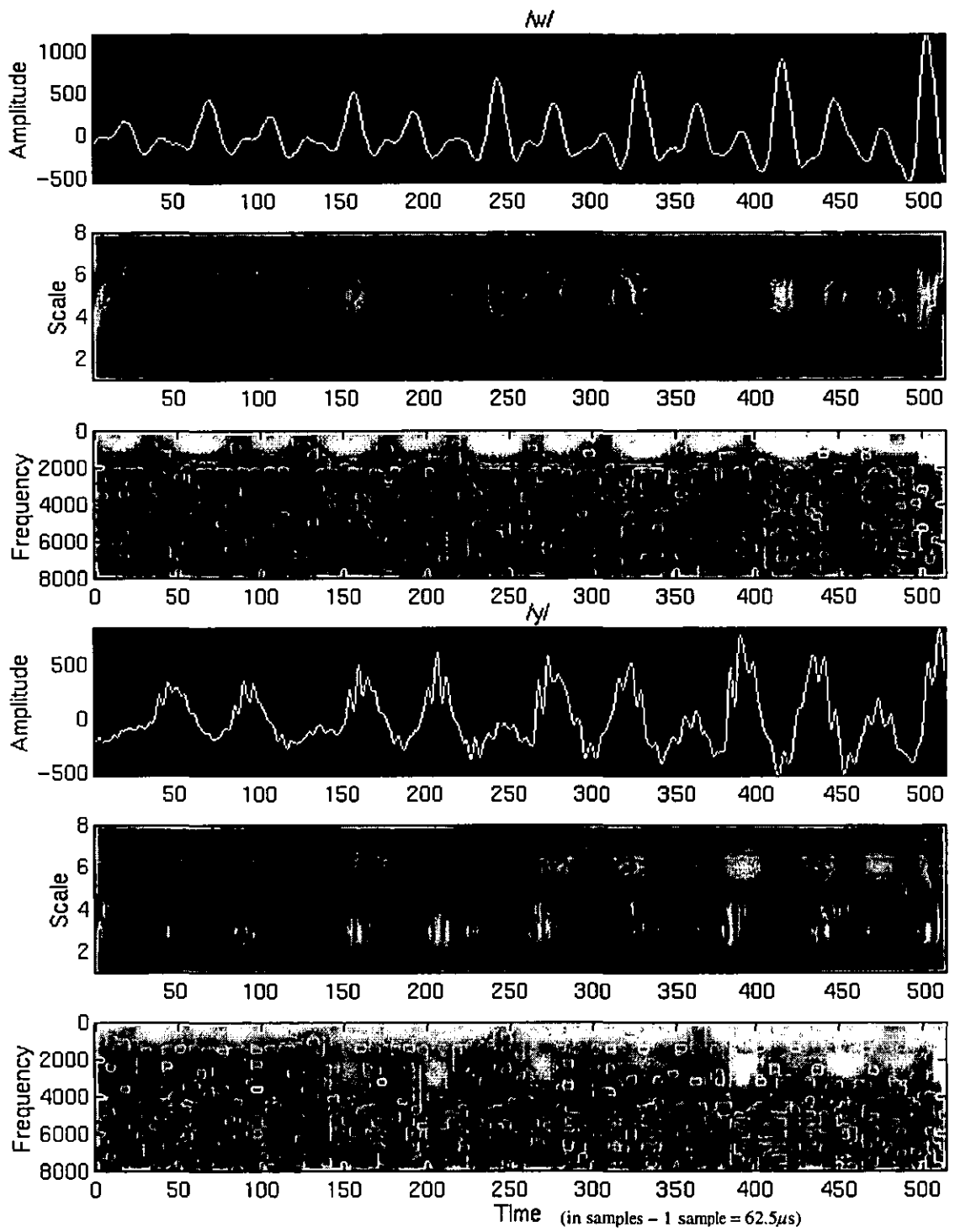
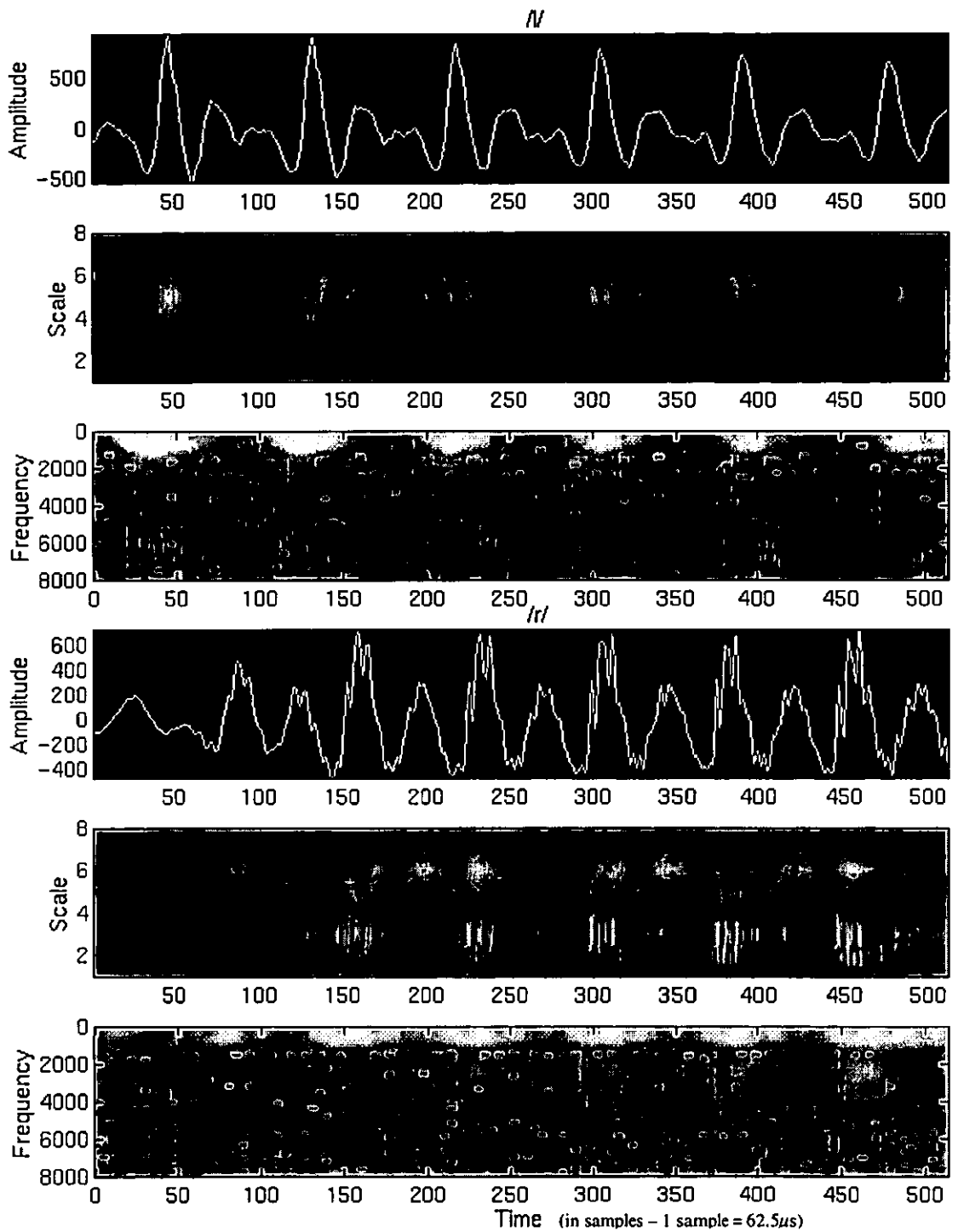


Figure 3.11 (a)



(b)

Figure 3.11 (a) and (b): Four typical examples of liquids and glides as used in Example 3.5. Note the periodic structure although they do have characteristics that change in the long term.

Technique		MisClassification Rate (%)
STFT_64	Tr	70.84
	Pr	72.26
DD3_64	Pr	61.18
	Tr	62.21

Table 3.7 Showing standard DWT algorithm overall misclassification on the 16 stressed vowels from dialect regions 1 and 2 of the TIMIT database using the DD3 wavelet.

		Predicted Class (Mahalanobis derived distances off-diagonal, %error on diagonal)															
True Class		iy	ih	eh	ey	ae	aa	aw	ay	ah	ao	oy	ow	uh	uw	ux	er
	iy	37.63	3.58	10.99	3.81	12.88	21.41	19.27	15.99	15.12	20.69	12.48	14.71	8.87	7.89	2.04	15.20
	ih		67.38	3.11	1.06	5.85	12.5	11.16	7.86	6.98	12.02	5.16	7.50	3.52	6.22	3.40	6.28
	eh			73.78	3.44	1.35	5.35	4.68	2.15	2.40	6.34	2.69	4.75	4.72	10.05	10.04	5.03
	ey				57.14	5.19	13.22	12.00	8.14	8.22	12.71	5.69	8.61	5.67	9.45	5.34	9.47
	ae					61.69	4.70	4.10	1.46	3.90	6.98	4.76	7.17	8.51	13.98	12.89	9.38
	aa						65.80	0.36	1.30	1.81	2.48	4.82	4.66	8.44	13.14	19.00	9.19
	aw							79.82	1.16	1.60	2.56	4.62	4.29	7.71	11.98	17.21	8.83
	ay								80.95	1.64	4.02	3.73	4.92	7.05	12.34	14.49	7.54
	ah									72.26	2.61	2.24	2.21	3.95	8.39	13.07	5.44
	ao										57.24	2.95	1.38	6.13	10.57	18.95	10.52
	oy											91.87	1.23	2.59	6.47	10.83	5.56
	ow												65.77	2.53	5.98	12.85	7.15
	uh													79.77	2.90	6.46	4.75
	uw														54.67	4.4	6.93
	ux															50.81	10.00
	er																40.22

Table 3.8: Showing standard DWT algorithm misclassification on the 16 stressed vowels from dialect regions 1 and 2 of the TIMIT database using the DD3 wavelet. The off-diagonal elements represent the Mahalanobis-derived distances from the top 64 wavelet features between classes that was used to calculate the final prediction rates. These can be seen to correspond well to Figure 3.4. For example /iy/ has the greatest distance from /aa/ both in terms of where it is articulated and its Mahalanobis distance (21.41). The same phone is seen to be close to /ih/ in both contexts also.

3.7 REFERENCES

- [1] Buckheit, J. and Donoho, D.L., "Improved Linear Discrimination using Time-Frequency Dictionaries," Technical Report Stanford University, 1995.
- [2] Buckheit, J.B. and Donoho, D.L., "WaveLab and Reproducible Research," ,” in *Wavelets and Statistics*, Lecture notes in statistics, ed. Antoniadis and Oppenheim, Springer-Verlag, 1995, pp. 55-81.
- [3] Cohen, L. "Time-Frequency Distribution – A Review," *Proc. IEEE*. Vol. 77, No. 7, pp. 941-981. 1989.
- [4] Coifman, R.R., Donoho, D.L., "Translation-invariant denoising," in *Wavelets and Statistics*, Lecture notes in statistics, ed. Antoniadis and Oppenheim, Springer-Verlag, 1995, pp. 125-150.
- [5] Daubechies, I., "The Wavelet Transform, Time – Frequency Localization and Signal Analysis," *IEEE Trans. on Inf. Th.*, vol. 36, No. 5, September 1990.
- [6] Donoho, D.L., "Interpolating wavelet transforms," *Technical Report 408*, Dept. Statistics, Stanford University, Stanford, CA, Oct. 1992.
- [7] Elbrond Jensen,H., Hoholdt, T., and Justesen,T. "Double series representation of bounded signals," *IEEE Trans. Inform. Theory*, vol. 34, pp. 613-624, 1988.
- [8] Gabor, D. "Theory of Communication," *J.IEE*, 93:429-457, 1946.
- [9] Holschneider, M., Kronland-Martinet, R., Morlet, J. and Tchamitchian, P., "A Real Time Algorithm for Signal Analysis with the Help of the Wavelet Transform," *Wavelets, Time-Frequency Methods and Phase Space*, pp. 289-297. Springer-Verlag, Berlin, 1989.
- [10] Jaffard, S. "Pointwise Smoothness, Two-Microlocalisation and Wavelet Coefficients," *Publicacions Matematiques*, 35:155-168, 1991.

- [11] Kadambe, S., "The Application of Time-Frequency and Time-Scale Representations in Speech Analysis," *Ph.D. Thesis*, Univ. of Rhode Island, Dept. of Electrical Engineering, 1991.
- [12] Kadambe, S. and Boudreaux-Bartels, G.F. "Application of the Wavelet Transform for Pitch Detection of Speech Signals, " *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [13] Kadambe, S and Srinivasan, P. "Applications of adaptive wavelets for speech," *Optical Engineering* 33(7), pp.2204-2211 (July 1994).
- [14] Mallat, S. "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 674-693, 1989.
- [15] Mallat, S. and Hwang, W.L., "Singularity detection and processing with wavelets," *IEEE Trans. Info. Theory*, 38(2):617-643, March 1992.
- [16] Mallat, S. and Zhong, S., "Characterisation of signals from multiscale edges, " *IEEE Trans. Patt. Recog. and Mach. Intell.*, vol. 14(7), pp.710-732, July 1992.
- [17] Mallat, S., "A Wavelet Tour of Signal Processing." *Academic Press*, 1998.
- [18] O'Shaughnessy, D. "Speech Communication: Human and Machine." New York: Addison-Wesley, 1987.
- [19] Rioul, O. and Vetterli, M., "Wavelets and Signal Processing," in *IEEE Sig. Proc. Mag.*, 8(4):14-38, October 1991.
- [20] TIMIT Acoustic-Phonetic Continuous Speech Corpus. National Institute of Standards and Technology Speech Disc 1-1.1, October 1990. NTIS Order No. PB91-505065.

Chapter Four

The Best Basis Algorithm for Phoneme Classification

4.1 Introduction

In order to provide a more flexible approach for representing the speech signals encountered in the previous chapter, our attention is turned to an adaptive wavelet modelling approach which uses the Best-Basis algorithm of Coifman and Wickerhauser [6]. This combines the advantages of using compactly-supported wavelet bases with the ability to select an adaptive tiling of the time-scale plane, providing a time-scale analysis dependent on the problem at hand.

The dyadic wavelet transform described in Section 3.3 partitions the time frequency plane as shown in Figure 3.2. It was noted that this type of time-frequency decomposition is usually referred to as a constant-Q or octave band decomposition. While the Heisenberg inequality principle still holds, it is clear that a given signal is analysed such that the frequency is partitioned in dyadic intervals whose support decays logarithmically with the frequency. By generalising this concept, Coifman et al [9] have introduced new families of dyadic orthonormal wavelets called *wavelet*

packets which decompose the frequency axis arbitrarily. These packets are translated uniformly in time to ensure the entire time-frequency plane is covered.

Orthonormal bases of $L^2(\mathcal{R})$ can also be constructed to do the opposite; i.e. basis functions can be constructed that partition the time axis in such a way as to guarantee that any disjoint interval on the real axis will be partitioned smoothly. These types of basis function essentially perform local Fourier analysis on the signal within the interval and include some signal values from the adjacent intervals. As a result, this type of transform is often termed the Local Trigonometric Transform.

This chapter will also consider the addition of these two types of basis function to a fixed yet flexible set of functions referred to as a *library* with which to analyse the speech data. A library is said to consist of a number of *dictionaries*. These dictionaries consist of a collection of waveforms ψ_t such that the dictionary is given by $D = (\psi_t : t \in \lambda)$, where the possibilities for λ are given below. Assuming our signals are of dyadic length $n(=2^J)$, where J is the maximum depth of decomposition, then a dictionary is said to be *complete* if it contains exactly n linearly independent wavelet bases. There will thus be a unique representation of a given signal in this dictionary. Typically the dictionaries used in this thesis are *overcomplete* or *redundant* and dynamic programming based searches are invoked to find the orthonormal representation of the signal. The parameter t signifies either:

- (i) Frequency.
- (ii) Time and scale jointly.
- (iii) Time and frequency jointly.

In this chapter, the third case will be considered where the dictionaries are wavelet and trigonometric packet dictionaries. This follows the philosophy of the *Best-Basis paradigm* outlined in [6] and [15]. The thrust of subsequent work will be to adapt this paradigm for classification.

Finally, the wavelet based feature extraction system will be outlined and its performance on phonetic classification problems similar to those described in the previous chapter given.

4.2 Why Wavelet/Trigonometric Packets?

The wavelet packet transform is a generalisation of the wavelet transform. For signals which are oscillatory in nature, such as speech, a transform like the DWT which partitions the frequency axis finely towards the lower frequencies may not be the most suitable choice. Furthermore, since speech signals vary so widely in nature, e.g. from the slowly varying, well behaved vowel sounds to the highly transitory plosive subclasses, a feature extractor capable of adapting its resolution and/or basis function accordingly would seem desirable. Some work has been done on applications in speech, e.g. [11] in which wavelet packet transforms were used to parameterise the audio signal prior to speech classification, or [17] in which suitable features were extracted for recognition purposes. In many cases, the aim of using wavelet or trigonometric packets is to achieve low bit-rate compression of speech signals for which these kind of techniques are well suited, e.g. [16]. Work which has used wavelet and trigonometric packets for feature extraction purposes such as in [11] seems to focus more on the direct substitution of this method into standard speech recognisers, however adaptation of the technique has concentrated more on real time implementation rather than on the extraction of suitable features for classification. This chapter concentrates on the quality of recognition features, and attempts to utilise the inherent flexibility of this method for phoneme classification. There are two main reasons for pursuing this approach. The first has to do with the two types of adaptivity in the transform; the selection of the basis function and the selection of the time-frequency tiling. The latter of these relates to a non-linear approximation, more suitable for signals that are not necessarily uniform or smooth, i.e. those that belong in *Besov* space. Wavelet packet approximations can be shown to yield superior results over conventional techniques e.g. Fourier and Karhunen-Loeve Transforms (KLT's), for these types of signals. Since Besov spaces are defined for signals that are *piecewise smooth*, they are generally considered a better model for real world signals such as speech than say *Sobolev* spaces which are spaces containing smoother signals. In fact the literal Russian translation of Besov means devilish. This concept is likely to be of use to speech processors who commonly encounter badly behaved complex signals which are usually (inefficiently) dealt with using techniques (Fourier, KLT) originally designed for Sobolev types of signal.

The second issue deals with how the best-basis is chosen. In the conventional best-basis paradigm, one chooses the best approximation from the wavelet/trigonometric packet table via some kind of cost function. This opens up a number of application-specific possibilities, for example if the aim is compression then least distortion in the Shannon entropy sense will be the final goal and, for example, a rate distortion measure could be used as a criterion for selection. Chapter 5 has incorporated recognition criterion into the selection rule to increase overall performance which in this case was the misclassification rate.

The following two subsections define the two types of Time-Frequency dictionaries used in the experiments.

4.2.1 Wavelet Packet Bases

Section 3.3.1 defined a multiresolutional space V_j which could be decomposed into a lower resolutional space V_{j+1} plus a detail space W_{j+1} by splitting the orthogonal basis $2^{-j/2} \phi(2^{-j}k - n)$, $n=1,2,\dots$ of V_j into two new orthogonal bases i.e.

$$2^{-(j+1)/2} \phi(2^{-(j+1)}k - n), \quad n=1,2,\dots \text{ for } V_{j+1}$$

and

$$2^{-j/2} \psi(2^{-j}k - n), \quad n=1,2,\dots \text{ for } W_{j+1}$$

Instead of iterating the process only on the low frequency band as in the standard dyadic wavelet transform, it can be shown [9] that this result can be generalised to allow division of the high frequency bands, thus deriving new bases. This results in a binary tree as shown in Figure 4.1. Each subspace in the tree is indexed by its depth j and number of subspaces p directly below it. The two wavelet packet orthogonal bases at a parent node (j,p) are defined by

$$\psi_{j+1}^{2p}(k) = \sum_{n=-\infty}^{\infty} h[n] \psi_j^p(k - 2^j n) \quad (4.1)$$

and

$$\psi_{j+1}^{2p+1}(k) = \sum_{n=-\infty}^{+\infty} g[n]\psi_j^p(k - 2^j n) \quad (4.2)$$

Examples of wavelet packet bases derived from the Daubechies wavelet are shown in Figure 4.2.

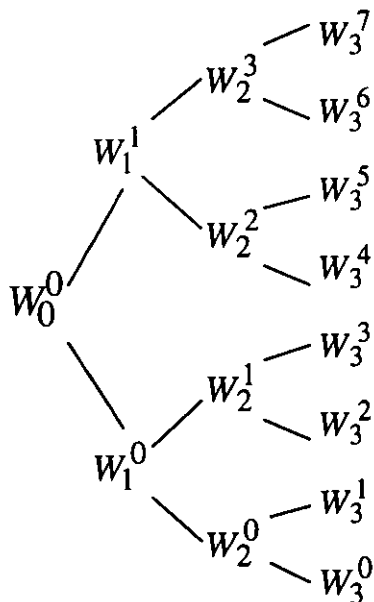


Figure 4.1: Binary tree showing the wavelet / trigonometric packet subspace decomposition. Once a signal is decomposed into a tree like this, the subspaces can be pruned according to their cost, resulting in an orthonormal basis.

The problem now is that this procedure results in an overcomplete basis. In fact in a full wavelet packet binary tree of depth J there are over 2^{2^J-1} different orthonormal bases to choose from. The subsequent aim, following decomposition into this tree (this is sometimes called a *packet table*) is to select the *basis* best suited for the problem at hand. This procedure is described in more detail in Section 4.3. Typically, as in the “Best” Basis algorithm, this is done adaptively depending on the signal to minimise an *additive* cost function. A common example of such a cost is the Shannon Entropy, which, along with the importance of additivity will be defined in Section 4.3. Figure 4.3 illustrates the Best-Basis methodology on part of the word ‘greasy’. Note the adaptive tiling of the time-frequency plane.

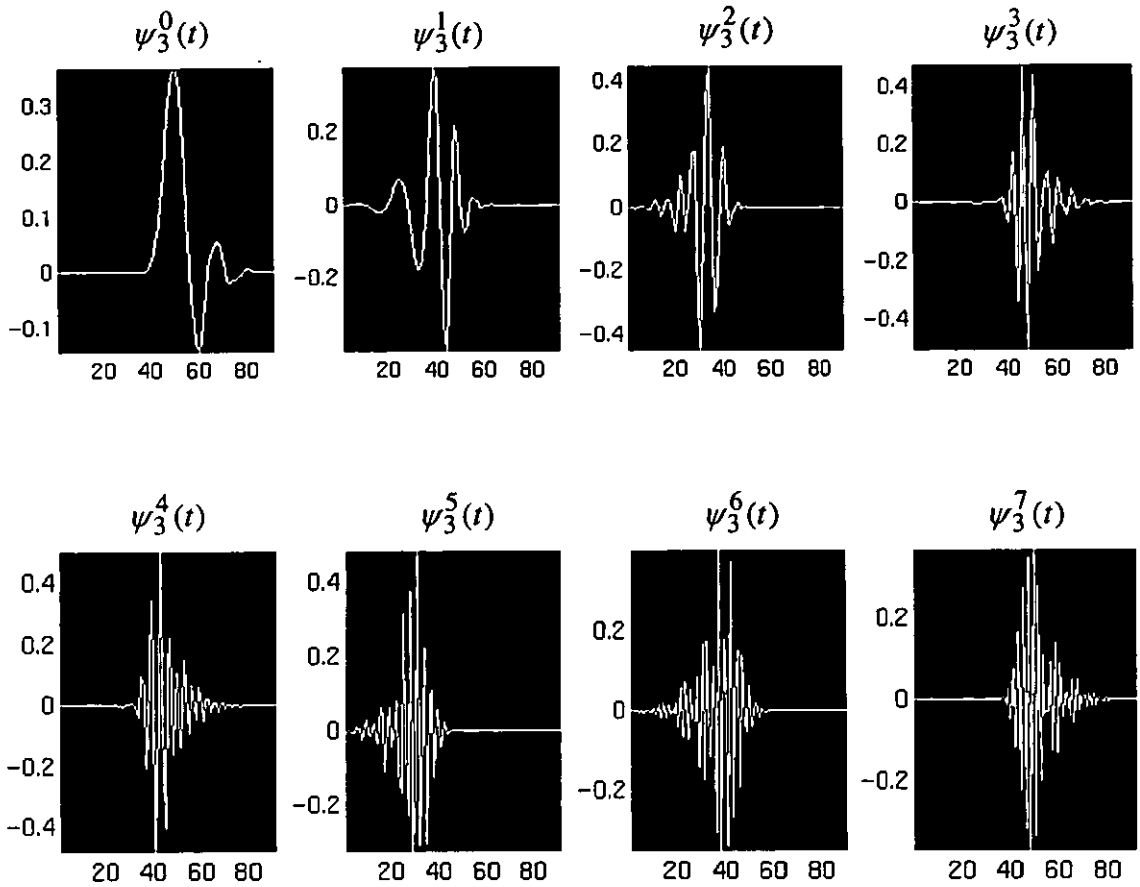


Figure 4.2: Some Wavelet Packets generated from the Daubechies 6 filter at the depth $j = 3$ of the binary tree. They are ranked in frequency order corresponding to their position in the binary tree. Note the much more oscillatory nature of Wavelet Packets compared to ordinary wavelets.

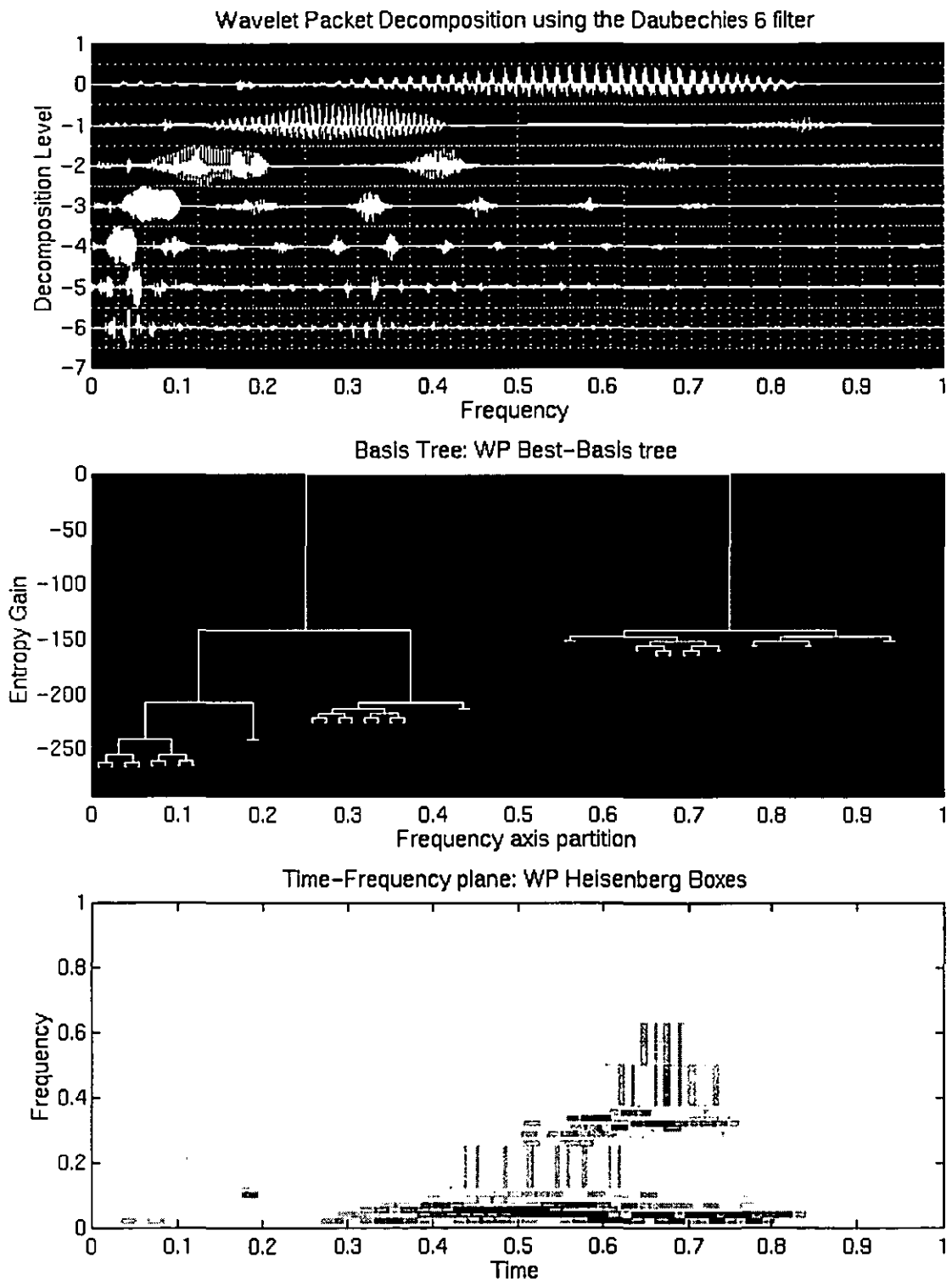


Figure 4.3 Illustrating Wavelet Packet decomposition on the 'grea' part of the word 'greasy'.

4.2.2 Lapped Orthogonal Transforms

It has already been seen that the wavelet packet transform, when used with a Best Basis search results in partitions of the frequency axis which are well adapted to the frequency content of the signal in that interval. Wavelet packets are thus very suitable for signals whose behaviour varies in frequency. If the function is more likely to be non-stationary, then it may be more useful to have a conjugate of the wavelet packet, sometimes called a *Block Transform* which is capable of partitioning the time axis according to the spectra contained in each interval. One of the consequences of the Balian-Low theorem mentioned in Section 3.3.1, has to do with short-time methods such as the STFT which involve multiplication of the signal with a smooth window. This theorem states that for any τ_0 and ω_0 , no smooth window exists of compact support such that equation (3.3) forms an orthonormal basis of $L^2(\mathfrak{R})$. However, Coifman and Meyer [7] proved that it is possible to partition the time axis into disjoint intervals smoothly by constructing overlapping basis functions on each interval. These functions are essentially a sinusoid, modulated by a cut-off function which has particular properties e.g. it must be even and have vanishing derivatives at its end points. One such window is the ‘bell’ function:

$$\beta(t) = \begin{cases} \sin \frac{\pi}{4} (1 + \sin \pi t) & \text{if } -\frac{1}{2} < t < \frac{3}{2} \\ 0 & \text{otherwise} \end{cases} \quad (4.3)$$

This window is composed of two even lapped projectors (half a window) which remove problems due to discontinuities encountered when using conventional orthogonal projectors which have more abrupt cut-offs. One projector is a raised smooth profile and the second is a decaying one. The window itself is generally defined on an interval; the method used in this thesis is exactly the same as the Discrete Cosine IV Transform [17] with the exception of this window function whose interval on the time axis is $I = [a_j, a_{j+1}]$, such that the window interval is positive,

bounded i.e. $a_{j+1} - a_j > \sigma$, and has a smooth cut-off, [1], [7]. The bell function, moreover, can be characterised by its cut-off response.

$$b_j(t) = \begin{cases} \beta \left(\frac{t - a_j}{r} \right) & (a_j - r) \leq t \leq (a_j + r) \\ 1 & (a_j + r) \leq t \leq (a_{j+1} - r) \\ \beta \left(\frac{a_{j+1} - t}{r} \right) & (a_{j+1} - r) \leq t \leq (a_{j+1} + r) \\ 0 & \text{elsewhere} \end{cases} \quad (4.4)$$

with $0 < r \leq \sigma$.

Via this window it is now possible to gain a new family of orthonormal bases by means of the following basis functions:

$$\varphi_{j,k}(t) = \sqrt{\frac{2}{|I_j|}} b_j(t) \cos \frac{2k+1}{2} \frac{\pi}{|I_j|} (t - a_j) \quad (4.5)$$

These functions are well localised in time. Also these functions have energy that is well-localised in the frequency domain. Figure 4.4 shows some examples of cosine packets in the time domain and Figure 4.5 illustrates the equivalent Best-Basis decomposition using cosine packets instead of the wavelet packets in Figure 4.3. Note that it is the time axis which is partitioned in this case.

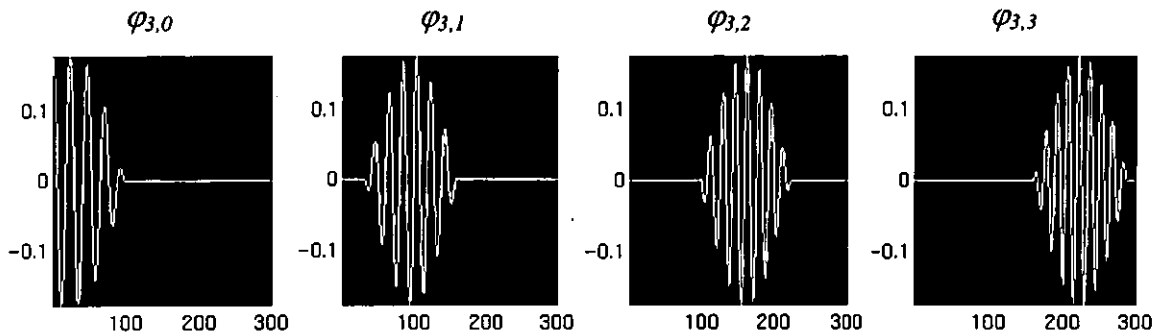


Figure 4.4 : Some Cosine Packets

In practice, this transform can be implemented by calculating the DCT-IV transform as usual after applying a preliminary folding step as described in [7]. This folding step essentially causes the overlapping parts of the bell to be folded inside the interval. If one performs this operation on the original signal $x(t)$ it becomes a series of disjoint signals, which, when the DCT-IV is applied, corresponds to taking the inner product:

$$\langle x(t), \varphi_{j,k}(t) \rangle.$$

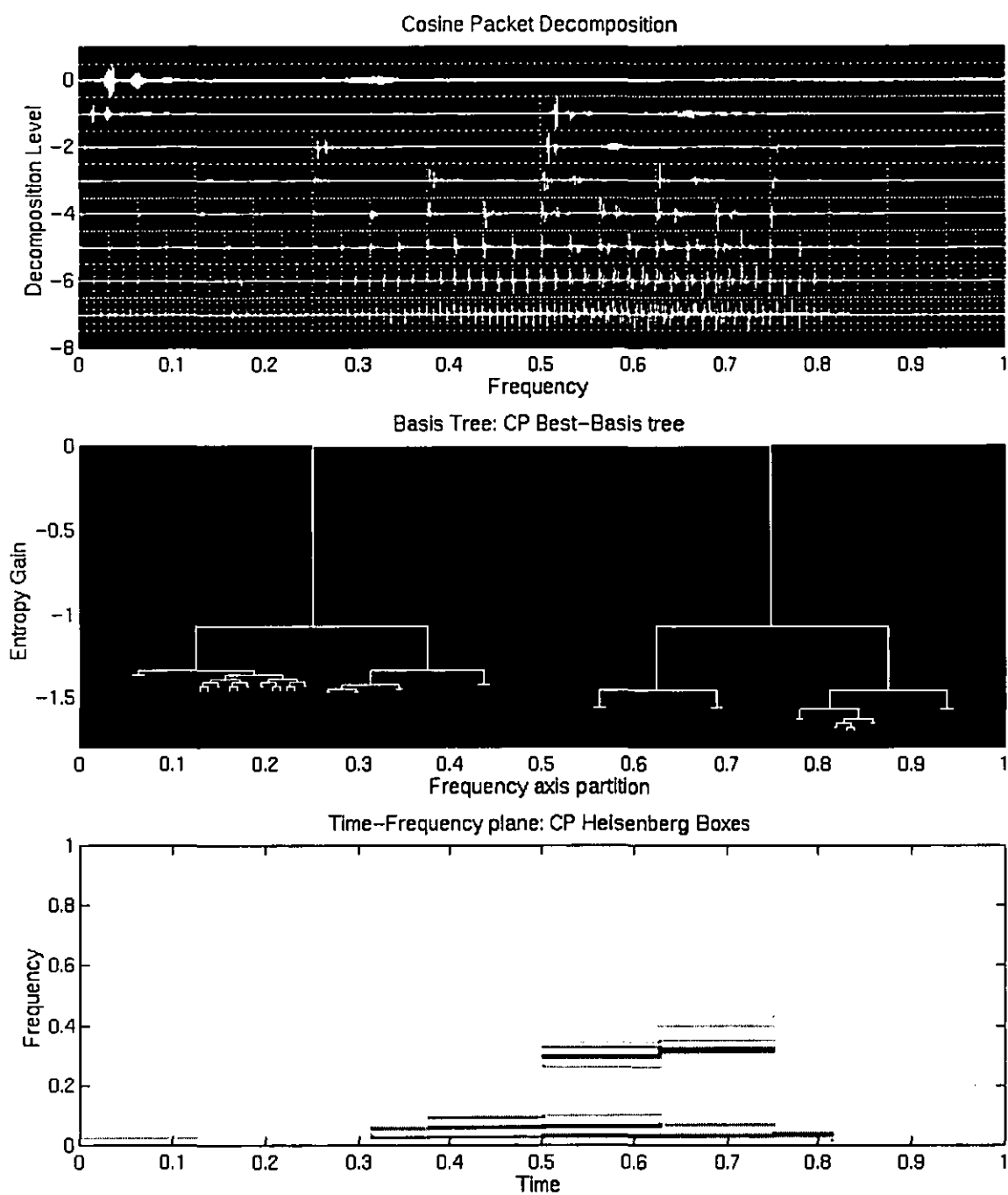


Figure 4.5: Illustrating Cosine Packet methodology on the speech segment of Figure 4.3.

4.2.3 Translation Invariance

Both Wavelet and Cosine Packet bases suffer from translation invariance similar to that exhibited by the DWT (Section 3.3.1). In both cases (i.e. wavelet and cosine packets), problems from the point of view of pattern recognition are likely to occur due to misalignment between signal features and basis function features although for subtly different reasons. While wavelet packet decompositions experience translation distortions for similar reasons to the DWT, the sine/cosine packet bases in equation (4.5) are sensitive to shifts in time. Methods, as mentioned in Section 3.3.1, also exist for wavelet and cosine packet transformations that can adjust for time and/or scale translations (see e.g. [3],[4],[8],[10] for various approaches and applications).

As the following section will show, the idea of transforming a signal into a wavelet packet transform amounts to a redundant representation with a data structure conforming to a balanced binary tree. When searched and pruned using some kind of application-dependent cost function, the representation becomes orthonormal. Clearly, an orthogonal transform for perfect reconstruction is necessary, and also, computational complexity apart, if one knows *a priori* that a transform is orthogonal, this can allow further processing in this domain such as thresholding for compression or denoising without violating important signal structures. One might question the importance of orthonormality for uses in classification; indeed one of the main approaches for dealing with shift-invariance is to provide an oversampled representation which is no longer orthonormal. However, it should be noted that such redundant transformations necessarily render subsequent interpretation of the expansion coefficients more difficult.

The effect of having two different wavelet transformations for the same signal within a time shift τ as described in Section 3.3.1 is also true for wavelet packet transforms. This artifact adversely influences the cost function which is calculated at each node in the tree prior to pruning.

This in turn results in a different set of subspaces ‘best bases’ for the same signal. Figures 4.6 and 4.7 clearly show this effect. The same is also true for trigonometric transforms.

The following set of experiments are based on the so-called ‘spin-cycle’ described in [8] which was originally developed for the purpose of providing translation-invariant denoising. In that setting, the problem of translation invariance manifested itself as a pseudo-Gibbs effect in the neighbourhood of discontinuities for wavelet-based denoising. In the case of wavelet and cosine packets, problems also arise in the region of segmentation points. Here are the steps modified slightly for the problem of classification:

Choose a range of shifts H and circularly shift each signal in the training set by $-h, -h+1, -h+2, \dots, -1, 1, \dots, h$ where $h \in H$, resulting in a total of $(2h+1) \cdot N^{(c)}$ signals (where $N^{(c)}$ is the number of signals belonging to a class (c)), including the originals, for each class in both training dataset. Naturally, the $(2h+1)$ signals have the same class assignment as the original. As usual, the best-basis for each signal is built and a classifier is designed on the top k -features. The test set is then fed into the classifier to give the prediction performance of the problem.

4.3 Selection of the ‘Best- Basis’

The organisation of the packet transform which is gained by recursively refining the input signal by segmentation of the time/frequency axis depending on whether trigonometric or wavelet packet transforms were used, results in a data-structure resembling a balanced binary tree. The standard dyadic wavelet basis used in Figure 3.2 exhibits the time-frequency tiling gained by sampling the continuous version on a dyadic grid. If one were to desire a specific time-frequency tiling, i.e. one which was adapted to a particular signal (or signal class), this would correspond to an irregular sampling grid locally adapted to the sharpness in the signal variations. This can be achieved in practice by pruning the binary tree in a manner that will minimise (or maximise) a given cost function. The cost is highly dependent on the application. For

compression, a measure that minimises distortion, e.g. Shannon entropy, would be a good choice, for classification a cost functional that best reflects the distance

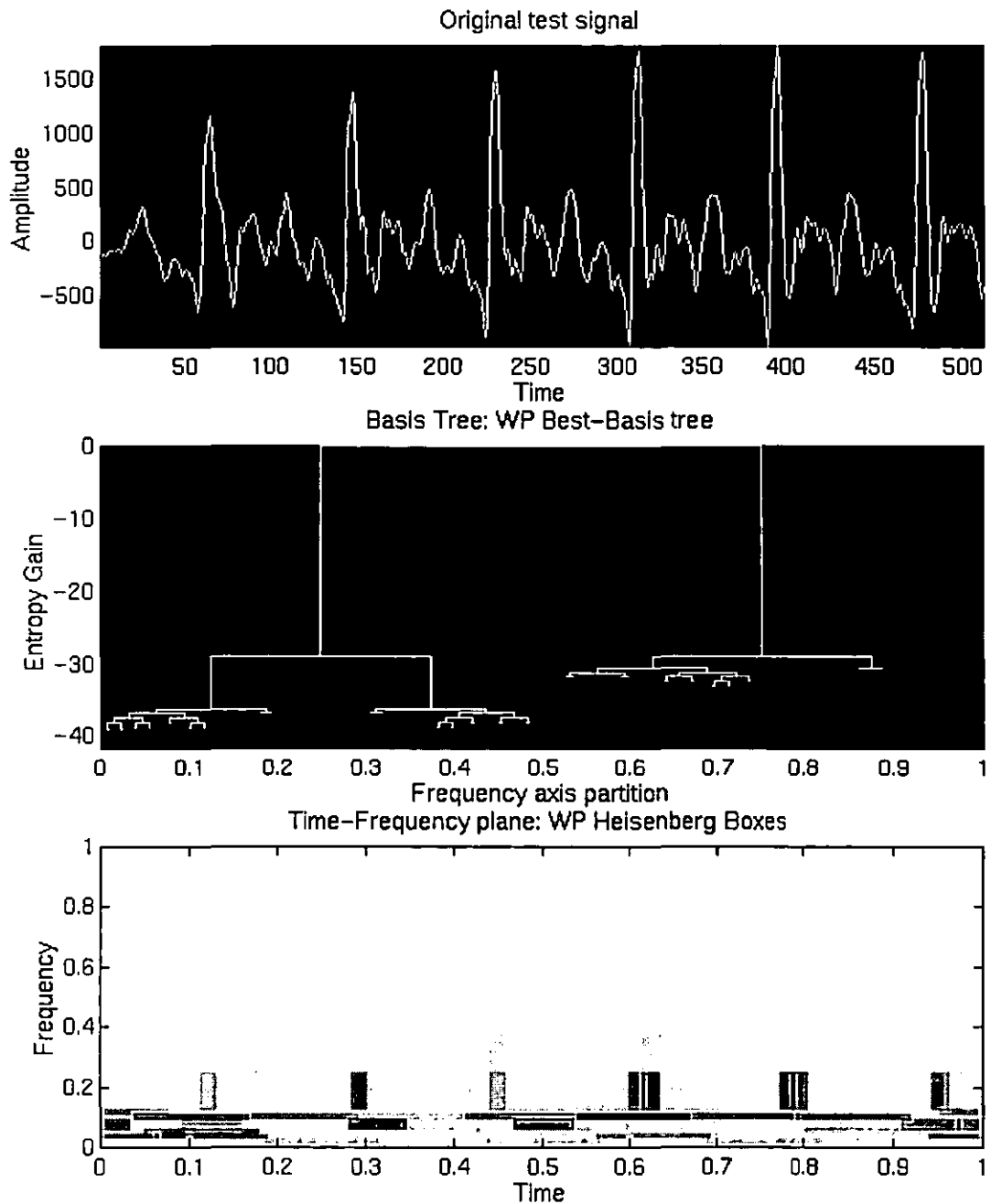


Figure 4.6 : Illustration of shift invariance in the wavelet packet transform on the /aa/ vowel sound. A different set of wavelet features are extracted from exactly the same signal that has undergone a time shift as shown in the next figure.

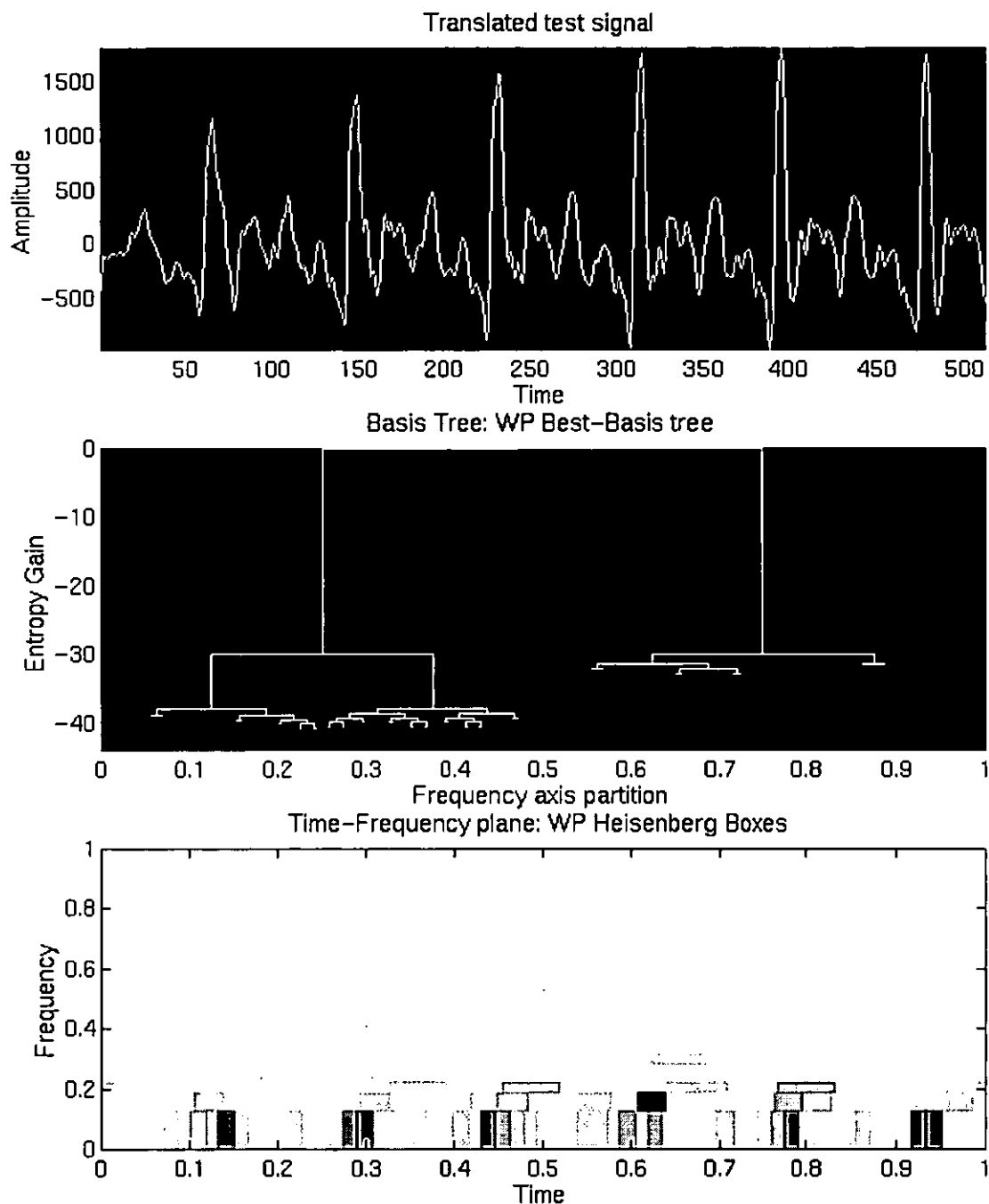


Figure 4.7 This plot shows the sensitivity of the wavelet packet transform on the signal in 4.6 after undergoing a 0.0625 ms time shift. One sees a quite different time-frequency structure. Note in particular the low frequency smooth tones corresponding to the underlying periodic structure of the vowel sound in Fig.4.6 are visibly reduced.

between the point clouds corresponding to different classes would be desirable. An example of this would be *relative* or *cross-entropy*. Here are some alternatives used in this chapter, first define \mathbf{p} as a non-negative sequence $\mathbf{p}=\{p_i\}$ where $\sum_i p_i=1$

Entropy is then defined as

$$H(\mathbf{p})=-\sum_i p_i \log_2 p_i \quad (4.6)$$

with the convention that $0.\log_2(0)=0$;

Define the l^2 based measure

$$l(\mathbf{p}) = \|\mathbf{p}\|_2^2 \quad (4.7)$$

i.e. the square of the l^2 norm. Note that this measure can be extended for all l^p space (with $p \geq 1$).

Define the logarithm of the energy as

$$M(\mathbf{p})=\sum_i \ln(p_i^2) \quad (4.8)$$

All costs used in this chapter have the property of *additivity* which allows a fast search with computational complexity $O(n \log n)$ for a wavelet packet dictionary and $O(n[\log_2(n)]^2)$ for a local trigonometric dictionary. This property is important because for a signal of dyadic length n , there exists a possible $2^{n/2}$ bases, i.e. the number of admissible choices of tree structure. To try and compute the best tree structure by straight comparison of all possible bases would cost about $n2^{n/2}$ operations which is prohibitive. In order to overcome this, the fast search Best Basis algorithm used in these experiments was introduced by Coifman and Wickerhauser [6] which also guarantees an orthonormal basis. A description of this method now follows.

First of all define \mathcal{G} as an additive cost functional and D as the dictionary into which the signals are to be expanded. The technique first takes a single signal and projects it into the chosen dictionary. The dictionary is then searched to minimise the chosen

cost function resulting in a collection of subspaces U_j^p spanned by the best basis functions A_j^p with $(j, p) \in \Lambda_M$ where Λ_M contains the indices of the M best subspaces/basis vectors. Also define B_j^p as the original set of basis functions belonging to the subspace U_j^p . The best set of A_j^p 's is found from the B_j^p 's by evaluating the cost of a parent subspace or node and comparing it with its two children nodes.

The algorithm is summarized as follows:

- (i) Initialise the best basis algorithm by deciding: (a) which dictionary (from a library) to use (i.e. to use wavelet packets or local trigonometric transforms); (b) what cost to use; (c) depth $J \leq \log_2(n)$ of tree into which the signals are decomposed.
- (ii) Expand signal x into chosen dictionary D obtaining a full packet table of expansion coefficients, $\{B_j^p x\}_{0 \leq j \leq J, 0 \leq p \leq 2^{j-1}}$.
- (iii) Set $A_j^p = B_j^p$.
- (iv) Determine the best subspace A_j^p for $0 \leq j \leq J, 0 \leq p \leq 2^{j-1}$ via the following

$$A_j^p = \begin{cases} B_j^p & \text{if } \mathcal{G}(B_j^p x) \leq \mathcal{G}(A_{j+1}^{2p} x + A_{j+1}^{2p+1} x), \\ A_{j+1}^{2p} \oplus A_{j+1}^{2p+1} & \text{otherwise} \end{cases}$$

- (v) Choose the top ($k < n$) features to train the classifier.

The last step is the so-called feature selection stage in speech recognition and is of paramount importance to the problem. Even the selection as to how many of k -features should be chosen is difficult question and is generally chosen empirically. In this chapter a rule of thumb was empirically chosen of ~10% of the total number of features available is used for training the data. The decision as to which features should be kept is based upon the expansion coefficients with the highest energy.

In Section 4.4 the regime described above is examined in the light of non-linear approximation and compared to classical linear methods which are seen to do best only under certain conditions.

4.4 Linear and Non-Linear Approximations in Bases

Section 3.3.1 examined the characterisation of function regularity in the Lipschitz sense. This was seen to encompass both global and local attributes of a function.

In this section, the wavelet transform is related to standard linear approximation techniques; it is noted that linear and non-linear approximations rely on different signal properties. When the functions are *uniformly smooth*, they can be characterised by their *Sobolev differentiability* which depends as the name suggests, on their differentiability. Sobolev functions can be shown to result in almost identical performance whether wavelet or Fourier methods are chosen. The performance criterion used is based on the minimum approximation error. However, when the functions have non-linearities themselves, they are best characterised in *Besov* space, which can be thought of as a model encompassing functions of Lipschitz and Sobolev regularity. Non-linear wavelet methods involving adaptive basis selection can be shown to provide greater speed of convergence for functions lying in Besov spaces compared with equivalent linear techniques.

4.4.1 The Linear Case

If a signal is uniformly smooth over an interval $[0,N]$ it may be precisely approximated by either a wavelet or Fourier basis. Fourier approximations are well suited to approximating uniformly smooth functions by using a partial sum of low frequency sinusoids. Measuring the local smoothness of a signal, furthermore, can be done by estimating the degree of its Sobolev differentiability. If $\hat{f}(\omega)$ represents the Fourier transform of $f(t)$; the Plancherel formula can be used to show that energy is conserved by the Fourier transform up to a factor of 2π

$$\int_{-\infty}^{+\infty} |f(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\hat{f}(\omega)|^2 d\omega$$

This notion can be extended for $f'(t)$, the first derivative of $f(t)$, for which the Fourier transform is $i\omega \hat{f}(\omega)$. Plancherel's formula proves that $f'(t)$ still lies in L^2 space if

$$\int_{-\infty}^{+\infty} |f'(t)|^2 dt = \frac{1}{2\pi} \int_{-\infty}^{+\infty} |\omega|^2 |\hat{f}(\omega)|^2 d\omega < +\infty \quad (4.9)$$

The function f is said to be *Sobolev differentiable* if

$$\int_{-\infty}^{+\infty} |\omega|^2 |\hat{f}(\omega)|^2 d\omega < +\infty \quad (4.10)$$

As in Section 3.3.1, the decay of the expansion coefficients is a measure of the

regularity; this decreases as ω gets large, typically like $\frac{1}{\omega^2}$.

In Section 3.3.1, it was seen that a function could be approximated by a Taylor series of polynomials. Likewise it can also be represented as a linear approximation. Taking an orthonormal basis of scaling functions $B = \{\phi_n\}_{n \in \mathbb{Z}}$, one can view the linear approximation problem for wavelets as follows

$$f_M = \sum_{n=0}^{M-1} \langle f, \phi_n \rangle \phi_n \quad (4.11)$$

The error of this approximation is thus

$$\varepsilon_M = \|f - f_M\|_2^2 \quad (4.12)$$

Since M is the number of basis vectors used in the approximation, it is desirable that the error ε_M decays rapidly toward zero as $M \rightarrow \infty$. This depends on two factors:

- (i) The type of basis used; Fourier, wavelet, or other e.g. the Karhunen-Loève transform.
- (ii) The nature of the signal. In the Fourier case, for example, a smooth signal will be well represented by the bottom few low frequency bases, whereas any localised irregularity such as a transient will require higher numbers of M to reduce ε_M ,

resulting in a slow decay. Truncation of the series, furthermore, leads to the appearance of Gibbs-like phenomena in the neighbourhood of discontinuities or other high frequency features.

Localised Fourier approximations of functions that are Sobolev differentiable can be achieved via Short-Time-Fourier-Transform which multiplying f with a window function prior to Fourier analysis. Section 3.3.1 showed that due to the Balian –Low theorem one is unable to have Fourier bases of simultaneous compact support and orthonormality. However, Section 4.2.2 gives the cosine functions of equation (4.5) as a set of orthonormal basis functions with compact support: one could certainly use these in the place of STFT. This in turn would be equivalent to taking the DCT-IV of the signals. Under such conditions, the performance of a localised Fourier method would then depend on the local signal regularity within the window.

Linear multiresolution approximations using wavelets amount to sampling on the grid of Figure 3.2. The accuracy in terms of mean square error approximation depends on the uniform smoothness of the signal. To derive the Sobolev differentiability m of a signal, a wavelet with enough vanishing moments $q > m$ is required to sufficiently describe the signal. For linear approximations of Sobolev signals on an interval, $f \in L^2[0, N]$ the periodic orthogonalised DWT behaves in a very similar way to a Fourier approximation, (see [14] for details). However, Sobolev spaces are not good models for real world signals, a much better approach assumes they are *piecewise smooth*. Such functions belong to *Besov* spaces, described in the following section.

4.4.2 The Non - Linear Case

Besov spaces are the space of functions more appropriate for real world signals where transients followed by smoother sections are often encountered. It has been shown in [14] that uniformly smooth Sobolev and uniformly Lipschitz α belong to Besov spaces. Not only does it provide a good model for these functions but also for functions that contain a finite number of transients. Figure 3.4 showed that wavelet coefficients were large in the neighbourhood of singularities but that there were relatively few of them. Keeping all but the largest $k < n$ is equivalent to constructing an adaptive time-frequency grid that is well-localised in the region of singularities.

Wavelet BB approximations are a refined version of the DWT which is still constrained to poor frequency resolution at high frequencies. Because the BB algorithm allows unconstrained tiling of the time-frequency plane with the Heisenberg boxes, it is a more adaptive non-linear approximation tool. Equation. (4.11) can now be written with the indices of the best M basis vectors contained in I_M :

$$f_M = \sum_{n \in I_M} \langle f, \phi_n \rangle \phi_n \quad (4.13)$$

The error of this approximation is

$$\varepsilon_M = \|f - f_M\|^2 = \sum_{n \in I_M} |\langle f, \phi_n \rangle|^2 \quad (4.14)$$

Therefore the linear approximation of equation (4.11) can be improved if the top (in the highest energy sense) M vectors are chosen according to the structures of f itself whether, using the BB algorithm, or by straightforward thresholding and ordering or by lastly using similar methods such as the matching pursuit [13]. The experimental section will attempt to ascertain under which circumstances, non-linear BB derived features provide improvement for the speech classification problems described in Chapter 3.

4.5 Results

Figure 4.8 shows the structure of the wavelet based acoustic-phonetic feature extractor used in the pre-classification stage for these experiments. The library of bases contained both wavelet packets and smooth localised cosine packets. The first stage of the system chooses the most suitable dictionary for the problem at hand. This can be done very simply in practice by choosing the one which gives minimum misclassification rate amongst them [5]. In a real world phonetic classification system, one could employ some kind of detector prior to this stage which could characterise into broad categories the type of signal coming in, e.g. voiced/unvoiced and choose the best dictionary based on *apriori* knowledge of the following results.

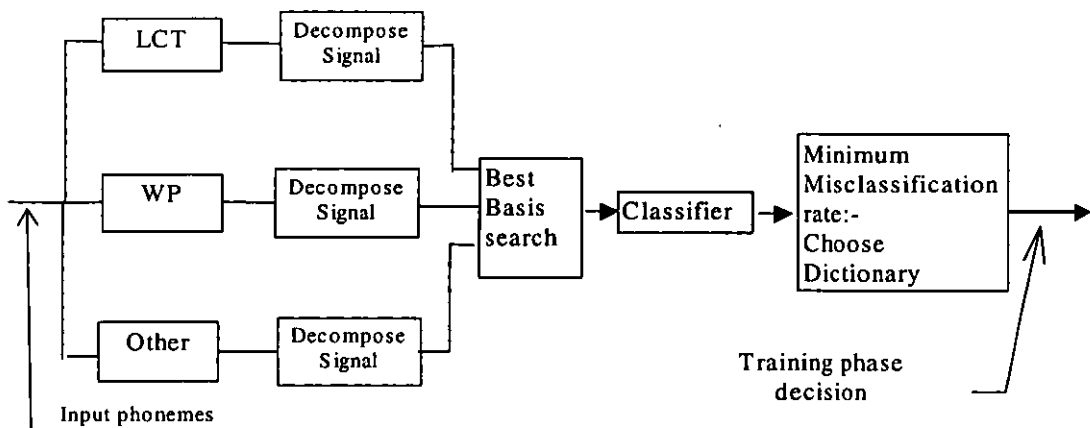


Figure 4.8: Training phase of proposed wavelet-based phonetic classifier. The final classifier is used to select the best dictionary for the given problem. Subsequent testing phases utilise the same dictionary.

The diagram illustrates the idea of choosing a dictionary of bases adaptively using the algorithm of Section 4.3. The choice of dictionary based on the least misclassification rate becomes part of the training phase for added flexibility.

In the following set of experiments, the aim is to determine the best dictionary for a given phoneme classification problem and to ascertain, furthermore the influence of the cost function used in basis selection on the final misclassification rate.

Using wavelet and cosine packets dictionaries, the Best Basis algorithm was implemented as described in Section 4.3 and applied to the speech data described in

Section 3.4. A single application-dependent wavelet basis function was chosen driven by the findings of Section 3.4 and from the basis, wavelet packets were generated via the relations (4.1) and (4.2). The cosine packet transform on the other hand has just one choice of basis function as shown in Figure 4.4. The five different cost functions used were (i) the direct sum of the coefficients within a node (ii) the l^1 measure of the subspace vector, equation (4.7), $p=1$ (iii) the l^2 measure at the node $p=2$, equation (4.7) (iv) the log of the energy, equation (4.8) (v) the Coifman – Wickerhauser entropy, equation (4.6).

4.6 Discussion

Table 4.2 (a) shows the performance of the Best Basis methodology on the phonemes /iy/, /aa/, /ax/ of Example 3.1. The cosine packet transform with the l^2 cost function provided the best misclassification rate (shown in bold) for this problem but does not perform as well as the DWT with the Deslauriers-Dubuc interpolating wavelet (see Table 3.3). Since the signal types analysed are modulated, and periodic, they are better analysed by the cosine packet. One reason why the DWT may perform better in this instance could be due to the packet transforms being too adaptive for the given problem. In other words, the DWT analyses all signals according to the same subspace decomposition (see Figure 3.3). Ideally in fact, for the purposes of classification, one would wish for a subspace decomposition that instead was well adapted to the characteristics of each class. In theory the Best Basis algorithm provides a tool with which to achieve, at least in part, this end; however, real world signals tend to contain noise and artifacts which vary from signal to signal and from class to class. The cost function in these cases has been shown in [12] to become a random variable, resulting for some cases in widely differing tree decompositions of acoustically similar signals belonging to the same class. This will clearly affect recognition performance

Table 4.3 shows the example of the plosive /p/,/t/,/k/ sounds. Coiflet wavelet packets with 10 vanishing moments perform the best, improving on the standard dyadic C_10 wavelet in Table 3.2 by about 6%. The major difference between performing a non-linear ordering of coefficients using Best Basis derived expansion coefficients and standard discrete wavelet methods lies not in the adaptive choice of sampling grid

which in both cases will be well localised in the region of discontinuities. Indeed the expansion coefficients in either case will be large throughout the cone of influence of rapidly varying sections. The improvement in performance is most likely due to the increase in adaptivity in the appropriate regions of the time-frequency plane. Unlike the DWT which has poor frequency localisation in the region of high frequencies, the BB algorithm has no such restriction except that the Heisenberg inequality principle still holds. For this kind of adaptive feature localisation, the BB algorithm is evidently better representing the internal correlation structure of these kinds of signals which are less regular and smooth than the voiced vowel sounds of the previous example and therefore better modelled by Besov spaces. These two disparate classification examples are probably the best illustrators of the comparative performance of the DWT and BB methods and the findings agree well with formal linear and non-linear signal approximation theory in bases. The remainder of the classification tasks show little difference to the DWT case. The results are summarised in Table 4.1 with some of the main results from Section 3.4 included for comparison. Tables 4.4 through Table 4.7 are included for completeness and contain the more detailed findings. It is seen that cosine packets do best in 4 out of the 6 cases tested. This is because cosine bases are oscillating in nature themselves and are thus well-suited to representing acoustic signals of similar characteristics. It is interesting to note that the wavelet packets provide improvement over both the cosine packets and the standard DWT in the two most difficult examples:- /p/, /t/, /k/ and /m/, /n/, /ng/ discrimination. Features enabling good classifier performance should provide good generalisation capabilities, giving low misclassification on unseen data also. The fact of relatively high misclassification in the training data of both these examples indicates the difficulty of the problem. Ultimately, the inherent flexibility of the wavelet packet approach in capturing the frequency structure of these signals via adaptive tiling is the likely cause of any improvement.

Table 4.8 shows the effect of removing variance in the wavelet transform due to translations of the signal via the Spin Cycle method. The two differing cases of /p/, /t/, /k/, Table 4.8 (a) and /iy/, /aa/, /ax/, Table 4.8 (b). It is seen that translation invariance only improves performance in the first case – this indicates that a feature for classification concerns the description of the transient rather than its position. For the more well-behaved type of signals in Table 4.8 (b), the Spin Cycle actually degrades

performance slightly implying that positional information is of some, but limited, importance for this problem.

Finally, regarding the issue of cost functions, it emerges that l^1 or entropy based costs are overall the safest choices for these kind of classification tasks. Indeed based on the work carried out in [2], time-varying structures are often best revealed by the l^1 cost.

Problem		Technique		Wavelet	Best Packet/ Cost
		DWT	BB		
/iy/, /aa/, /ax/	Tr	9.10	8.50	DD3	CP/ l^2
	Pr	7.55	8.39		
/p/, /t/, /k/	Tr	30.88	33.51	C10	C10/log
	Pr	42.26	35.80		
/m/, /n/, /ng/	Tr	41.80	41.36	D8	D8/ l^1
	Pr	46.40	45.02		
/f/, /T/, /s/	Tr	17.68	16.83	Haar	CP/Sum
	Pr	17.25	17.43		
/v/, /dh/, /z/	Tr	17.43	16.47	C10	CP/ l^1
	Pr	17.77	16.35		
/w/, /y/, /l/, /r/	Tr	20.36	21.99	Triangle	CP/ l^2
	Pr	24.33	25.70		

(Tr – Training Misclassification Rate Error %)

(Pr – Prediction Misclassification Rate Error %)

Table 4.1 Summary of best performance between the discrete wavelet transform, the wavelet packet and the cosine packets. See Table 3.1 for a list of keys. All results were gained using the top 64 coefficients out of a possible 512.

4.7 Summary

In this chapter, the performance of the Best Basis algorithm, originally designed for compression has been examined as a feature extractor for phonetic classification. The method chooses a suitable basis function and decomposes the time-frequency plane in an adaptive way. The top few coefficients (~10% of total), ranked by their energy, were chosen as suitable features for recognition

It is likely that the BB algorithm although providing a compact representation of the signals, is adapting extremely well to the internal structure of the speech. The

problems with this are potentially twofold. First, artifacts which in speech take the form of inter-speaker variation (e.g. rate at which a sound is uttered, variations in pitch, differing formant structures etc.). For this type of information, one would ideally like to discard or at least de-emphasise it for the purposes of classification. Using efficient non-linear well-localised wavelet features certainly yields a good estimate of the correlation in the sound and using the BB approach, one now has a method of dealing adaptively with the variant nature of the speech itself by using relations of the orthogonal bases of Chapter 3, i.e. wavelet and cosine packets of which the cosine packet was seen to perform best overall in comparison to wavelet packets. These bases have well-localised time-frequency characteristics. However the BB approach may be adaptive in the wrong sense, features for discrimination should try and capture the probability structure of the classes as a whole and provide the classifier with a good, again small set of features designed specifically for classification.

The final point regarding the lack of translational invariance exhibited by discrete wavelet methods concluded that this phenomenon was very much problem dependent. For example, if one is more concerned about the location of various features as information for recognition, the subtly different characterisations due to shift invariance are likely to matter little in the overall performance.

In the following chapter, ways of building discriminant attributes into the Best Basis paradigm will be examined.

Technique		Costs				
		Sum	$l^p(p=1)$	$l^p(p=2)$	log	Entropy
StdBBon	Tr	10.1	10.87	21.91	10.57	10.28
D6_64	Pr	11.61	10.49	28.67	10.91	9.93
StdBBon	Tr	10.40	10.40	8.5	30.12	10.4
CP_64	Pr	9.37	9.37	8.39	32.73	9.37

(Tr – Training Misclassification Rate Error %)

(Pr – Prediction Misclassification Rate Error %)

(a)

		Predicted Class		
		iy	aa	ax
True Class	iy	221	22	6
	aa	9	114	5
	ax	6	12	320

(b)

Table 4.2 Showing (a) Standard Best Basis algorithm on /iy/, /aa/, /ax/ dataset using different costs and basis functions. Splitting depth J was chosen as 6 for both dictionaries, and (b) LDA Confusion Matrix for testing datasets using Best-Basis Coefficients derived via the Cosine Packet transform with an l^2 cost.

Technique		Costs				
		Sum	$I^p(p=1)$	$I^p(p=2)$	log	Entropy
StdBBon	Tr	31.52	32.25	32.57	33.51	32.15
C10_64	Pr	43.23	37.75	42.26	35.80	38.06
StdBBon	Tr	30.58	30.47	35.18	49.01	30.58
CP_64	Pr	42.26	41.61	48.4	61.94	42.26

(Tr – Training Misclassification Rate Error %)
(Pr – Prediction Misclassification Rate Error %)

(a)

		Predicted Class		
		p	t	k
True Class	p	35	6	19
	t	16	85	21
	k	34	15	79

(b)

Table 4.3 (a) Showing Standard Best Basis algorithm on /p/,/t/,/k/ dataset using different costs and basis functions. Splitting Depth J was chosen to be 6. and (b) LDA Confusion Matrices for the testing datasets using the C10 wavelet packet with log-energy cost functional. The results show an overall improvement of around 6% compared to the best DWT result.

Technique		Costs				
		Sum	$l^p(p=1)$	$l^p(p=2)$	log	Entropy
StdBBon	Tr	41.87	41.36	41.25	41.84	41.36
D8_64	Pr	45.82	45.02	45.60	46.33	46.33
StdBBon	Tr	45.00	44.73	44.33	60.26	44.86
CP_64	Pr	47.78	47.70	46.47	62.62	47.49

(Tr – Training Misclassification Rate Error %)
(Pr – Prediction Misclassification Rate Error %)

(a)

		Predicted Class		
		m	n	ng
True Class	m	328	111	62
	n	187	348	173
	ng	33	53	80

(b)

Table 4.4 (a) Showing Standard Best Basis algorithm on /m/,/n/,/ng/ dataset using different costs and Basis functions. Splitting Depth J was chosen to be quite coarse at 4. for wavelet packets and 6 for cosine packets. and (b) LDA Confusion Matrices for the testing datasets using the D8 wavelet packet with an l^1 cost function.

Technique		Costs				
		Sum	$I^p(p=1)$	$I^p(p=2)$	log	Entropy
StdBBon	Tr	17.84	17.67	17.99	17.92	17.44
	Pr	18.08	17.86	17.86	18.35	18.61
D7_64	Tr	16.83	18.00	26.47	37.64	16.83
	Pr	17.43	17.91	27.23	36.74	17.42

(Tr - Training Misclassification Rate Error %)

(Pr - Prediction Misclassification Rate Error %)

(a)

		Predicted Class		
		f	T	s
True Class	f	399	149	35
	T	31	68	8
	s	83	90	1410

(b)

Table 4.5 (a) Showing Standard Best Basis algorithm on /f/,/T/,/s/ dataset using different costs and Basis functions. Splitting Depth J was chosen to be fairly coarse at 4. for wavelet packets and 6 for cosine packets. and (b) LDA Confusion Matrices for the testing datasets using the cosine packet transform with the basic Sum cost function.

Technique		Costs				
		Sum	$l^p(p=1)$	$l^p(p=2)$	log	Entropy
StdBBon	Tr	18.55	18.21	20.26	17.84	17.87
C10_64	Pr	19.09	17.11	21.83	18.71	17.39
StdBBon	Tr	16.58	16.47	21.86	34.62	16.64
CP_64	Pr	16.73	16.35	22.31	35.35	16.63

(Tr – Training Misclassification Rate Error %)

(Pr – Prediction Misclassification Rate Error %)

(a)

		Predicted Class		
		v	dh	z
True Class	v	195	61	8
	dh	51	110	9
	z	19	25	580

(b)

Table 4.6 (a) Showing Standard Best Basis algorithm on /v/,/dh/,/z/ dataset using different costs and Basis functions. Splitting Depth J was chosen at 6. for both transforms and (b) LDA Confusion Matrices for the testing datasets using the cosine packet transform with an l^1 cost function.

Technique		Costs				
		Sum	$l^p(p=1)$	$l^p(p=2)$	log	Entropy
StdBBon	Tr	23.08	23.06	41.87	22.40	22.23
	Pr	28.79	28.50	44.92	28.26	28.20
StdBBon	Tr	21.35	23.77	21.99	50.58	21.33
	Pr	26.35	28.68	25.70	54.73	25.94

(Tr – Training Misclassification Rate Error %)
(Pr – Prediction Misclassification Rate Error %)

(a)

		Predicted Class			
		w	y	l	r
True Class	w	204	20	52	41
	y	4	113	3	3
	l	97	35	455	77
	r	41	18	41	477

(b)

Table 4.7 (a) Showing Standard Best Basis algorithm on /w/,/y/,/l/,/r/ datasets using different costs and Basis functions. Splitting Depth J was chosen at 6. for both transforms. and (b) LDA Confusion Matrices for the testing datasets using the cosine packet transform with an l^2 cost function.

Technique		Block Length					
		3	4	5	6	7	8
StdBBon C10_64u l ¹ aSC	Tr	32.91	33.30	33.32	33.32	33.58	33.59
	Pr	37.96	37.57	37.80	37.80	37.51	37.14

(Tr – Training Misclassification Rate Error %)
(Pr – Prediction Misclassification Rate Error %)

(a)

Technique		Block Length					
		3	4	5	6	7	8
StdBBon CP_64u l ¹ aSC	Tr	10.53	10.69	10.75	10.91	10.79	10.99
	Pr	9.74	9.72	9.93	10.19	10.19	10.23

(Tr – Training Misclassification Rate Error %)
(Pr – Prediction Misclassification Rate Error %)

(b)

Table 4.8 Illustration of the effect of the Spin-Cycle on (a) The /p/, /u/, /k/ problem using the Coiflet filter with 10 vanishing moments and the l¹ cost functional. It shows an improvement of 37.14% error on the prediction compared with 37.75% without translation invariance correction. (b) Is the /iy/, /aa/, /ax/ problem which indicates perhaps surprisingly a slight decrease in performance when Spin Cycle is used. The misclassification error rises to 9.72% compared with 9.37% before the procedure.

4.8 References

- [1] Auscher, P., Weiss, G., Wickerhauser, M.V. "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, 1992, pp. 237-256.
- [2] Chen, S., and Donoho, D.L., "On Basis Pursuit," *Technical Report (1994), Dept. of Statistics, Stanford Univ.* ftp://playfair.stanford.edu/pub/chen_s/asilomar.ps.Z.
- [3] Cohen, I., Raz, S., Malah, D., "Orthonormal Shift-Invariant Adaptive Local Trigonometric Decomposition," *Signal Processing*, vol. 57, No.1, Feb. 1997, pp.43-64.
- [4] Cohen, I., Raz, S., Malah, D., "Orthonormal Shift-Invariant Wavelet Packet Decomposition and Representation," *Signal Processing*, vol. 57, No.3, Mar. 1997, pp.251-270.
- [5] Coifman, R., Majid, F., "Adapted Waveform Analysis and Denoising." *Progress in Wavelet Analysis and Applications (Y. Meyer and S. Roques, eds.)*, Editions Frontieres , B.P.33, 91192 Gif-sur-Yvette Cedex, France, 1993, pp. 63-76.
- [6] Coifman, R.R. and Wickerhauser M. V. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [7] Coifman, R.R., and Meyer, Y. "Remarques sur l'analyse de fourier a fenetre." *Comptes Rendus Acad. Sci. Paris, Serie I* 312 (1991), 259-261.
- [8] Coifman, R.R., Donoho, D.L., "Translation-invariant denoising," *in Wavelets and Statistics, Lecture notes in statistics*, ed. Antoniadis and Oppenheim, Springer-Verlag, 1995, pp. 125-150.

- [9] Coifman, R.R., Meyer, Y., Wickerhauser M.V. "Wavelet Analysis and Signal Processing, " *In Wavelets and their applications*, pp. 153-178, Boston, 1992. Jones and Barlett. B. Ruskai et al eds.
- [10] Del Marco, S. "Time/scale adjusted dyadic wavelet packet bases." In Proc. SPIE vol.2762 pp 70-79, 1996.
- [11] Drygajlo, A. "New Fast Wavelet Packet Transform Algorithms for Frame Synchronized Speech Processing," *Proc. of the 4th Int. Conf. on Spoken Language Processing (ICSLP96)* (Philadelphia PA, USA 1996), vol.1, pp.268-271.
- [12] Krim, H., Pesquet, J.-C., "On the Statistics of Best Bases Criteria," *in Wavelets and Statistics*, Lecture notes in statistics, ed. Antoniadis and Oppenheim, Springer-Verlag, 1995, pp. 193-207.
- [13] Mallat, S., and Zhang, Z., "Matching Pursuits with Time-Frequency Dictionaries," *IEEE Trans. on Sig. Proc.*, 41(12):3397-3415, December 1993.
- [14] Meyer, Y., "Wavelets and Operators," *Advanced Mathematics*. Cambridge university press, 1992.
- [15] Saito, N. "Local feature extraction and its application using a library of bases." *Phd thesis*, Yale University (1994).
- [16] Thripraneni, J.A. et al, "Mixed Malvar Wavelets for Non-stationary Signal Representation," *Proc.ICASSP vol.1 pp.13-16, Atlanta* (1996).
- [17] Wesfried, E; Wickerhauser, M.V. "Adapted local trigonometric transforms and speech processing," *IEEE SP* 41, 3597-3600 (1993)
- [18] Wickerhauser, M.V. "INRIA Lectures on Wavelet Packet Algorithms." *INRIA*, Roquencourt, France, 1991. Minicourse lecture notes. Available on www at <http://wuarchive.wustl.edu/doc/techreports/wustl.edu/math/papers/inria300.ps.Z>.

Chapter Five

Discriminant Wavelet Bases and their Applications

5.1 Introduction

In this chapter, a new feature extraction methodology based on the Best-Basis algorithm [5] (the Local Discriminant Basis algorithm is examined, which has been designed specifically for the discrimination problem. The algorithm is modified by creating a closer tie between feature extraction and classification stages using a cost gained from the classifier from which to choose the best set of subspaces and subsequently, features.

A training phase is involved during which the final classifier is invoked to associate a cost function (a proxy for misclassification) with a given resolution. The sub-spaces are then searched and pruned to provide a Wavelet Basis best suited to the classification problem. Comparative results are given of the two methods illustrating their relative performances using the differing subclasses of speech considered in Chapters 3 and 4.

It has been seen in Chapters 3 and 4, that multi-resolution feature extraction is a useful way of representing non-stationary real world signals such as speech, see also [7], [13], [14].

Coupled with integrated optimisation of the feature extraction and classification stages, the aim is to provide an overall improvement in recognition performance. This problem is particularly important because as modelling techniques have improved vastly in recent years, further gains in recognition accuracy are likely to come from the pre-processing stage.

Wavelets and related techniques such as subband coding have been applied with considerable success to speech processing applications such as compression [18], [20], and to a more limited extent on feature extraction for speech recognition / classification [6], [11].

Their main advantages as seen in Chapters 3 and 4, are a somewhat richer multiresolution representations of the acoustic signal and the added flexibility of using one of a number of basis functions. Subsequent refinements that aim to efficiently model signal statistics by choosing the depth of resolution projection and amount of signal reduction adaptively [15] serve to improve accuracy of the model further.

Learning from the training set the best set of subspaces in which to model the data, results in a discriminant basis set which will highlight, using the expansion coefficients of the wavelet transform (preferably just a few), the major differences between classes. If feature reduction is subsequently carried out, then the final classifier (LDA) is designed in lower dimensional space. Assuming the data is well modelled in the first place, then there are a number of advantages to this approach. Overall performance, instead of being worsened is improved since the number of coefficients needed for discrimination is typically less than that required for representation tasks. This crucial *feature selection* stage both increases robustness and accuracy of the final classifier as well as reducing costly training times.

In this Chapter, an alternative implementation of this theoretical framework for phoneme classification problems is considered.

5.2 Problem Definition

In discriminant feature extraction, the aim is to project (somehow) signal classes onto the most statistically important components i.e. those which result in the best separability amongst the classes. Typically, the type of features extracted in this way give little insight to

the signals they are describing, however if chosen well, should increase the robustness of the classifier (e.g. see Figure 5.1). The questions one should ideally ask are (i) What optimality criterion should be used, i.e. how does one decide ‘what’s important’?

(ii) Given a particular decomposition, what features should be used upon which to design the final classifier.

With regard to (i) the criterion should relate to the particular problem, in other words, a measure of the separability between the signal classes should govern which projections are chosen. Point (ii) relates to the so-called ‘Curse of dimensionality’ encountered in statistical classification, i.e. how ‘reduction of dimensionality’ can be efficiently achieved for a given set of training classes.

These issues are not new; they form the backbone of discriminant statistical feature extraction design issues, what this chapter will show is the applicability of wavelet based methodologies which involve the pruning/growing of the wavelet/cosine packet binary tree given a speech classification (rather than representation) problem.

As in the case of the JBB (Join Best Basis) algorithm described in [19], analogies may also be drawn between this kind of approach and Principal Component Analysis (PCA) also known as the Karhunen-Loeve Transform (KLT) or looked at another way - Linear Discriminant Analysis (LDA). In both cases, BB and PCA, the signal is first decomposed onto its eigenvectors and then approximated using the most ‘statistically significant’ portion by choosing the top few eigenvectors depending on the task. For data representation the KLT is generally used, while for classification the choice would be LDA. In essence, the methods described in this chapter aim to ease the task of methods such as LDA which can’t deal too well with the direct input of high dimensionality input signals. The other main difference of the technique described here is that the top few vectors are chosen using a measure based on class separability. Furthermore, the basis functions are designed to give a concise representation and are able to analyse accurately the local time-frequency characteristics of the data – something LDA on its own is unable to do.

The process of classification can be described more formally in terms of a signal space (where all signals in the training and testing sets exist) called $X \subset \mathcal{R}^n$. Here the

dimensionality of each signal is n . The second space contains a list of the class names or *labels* which are assigned to each member of a particular class prior to training. This shall be called $T = \{1, 2, \dots, N\}$. Classification is then normally defined as a mapping between these two spaces. Feature extraction is then desirable for two main reasons: (i) As mentioned earlier, the ‘Curse of Dimensionality’ causes difficulty in accurately modelling a class due to redundancy and gives rise to larger computational costs depending on the classification method. (ii) Most real world signals like speech contain undesirable components which are unrelated to signal characteristics. These serve to confuse the classifier, reducing its ability to generalise; robustness of performance can be improved by a good feature extraction regime which excludes such components. To this end, a *feature space* $\mathfrak{F} \subset \mathcal{R}^k$; $k \leq n$ is set between the signal space and the class space such that a feature extractor f provides the map $f : X \rightarrow \mathfrak{F}$ and the final classifier g as providing the map $g : \mathfrak{F} \rightarrow T$. In this chapter, the performance of the whole system will be measured by the overall misclassification rate.

5.3 The Local Discriminant Basis Algorithm

This section recaps on the pioneering work described in [15] in which the LDB algorithm is first constructed. Recall that in Chapter 4, the Best Basis Algorithm was designed mainly for signal *compression* but that these features, if applied to speech, could still be considered good features for recognition since they address at least one of the aims of parametrization outlined in Section 5.2; they reduce the dimensionality of the signal by retaining those vectors that maximise the information content, in the Shannon Entropy sense, of the signal. This should have the effect of at least ensuring the Best-Basis features are robust and compact but are they really suitable for decomposing large numbers of signals as in the training sets dealt with here? The main advantage of an LDB type construction is that only one overall basis tree is calculated based on a proxy for the probability distributions of each class: the so-called time-frequency energy map defined in the following section. The decision process affecting the pruning of this tree is, instead of Shannon Entropy, based upon the relative discriminatory power of each subspace i.e. how much worth, from the separability point of view, is gained by descending a level further down the tree? Not only, therefore, does the LDB use a more intuitive type of criterion for gaining classification features, but it works out

the Best Basis – in the discriminant sense – only once. All subsequent signals in training and testing classes are subject to exactly the same decomposition. There is, therefore, a fairly costly training phase during which LDB is calculated but after that, the cost of the transform is comparable to the conventional Dyadic Wavelet Transform and subsequently the Short-Time-Fourier Transform.

Again, the LDB is capable of utilising a number of dictionaries of basis function, but in these experiments will be considered, as in Chapter 4 only the Wavelet and Cosine Packets.

5.3.1 Cost Measures

These cost functions, of which there are a number of different types, can be *additive*, in which case the pruning mechanism in Chapter 4 is *fast*, i.e. $O(LN)$ where $L = \lceil \log_2 N \rceil$. All essentially provide a measure of ‘energy concentration’ of the signal in question.

Definition 5.1. An additive cost function \mathcal{G}^{add} from a sequence $\{x_i\}$ to \mathfrak{R} is said to be additive if $\mathcal{G}^{add}(0)=0$ and $\mathcal{G}^{add}(\{x\}) = \sum_{i=1}^n \mathcal{G}^{add}(x_i)$.

Definition 5.2. The inequality $\mathcal{G}^{add}(z) \leq \mathcal{G}^{add}(x,y)$ between vectors $x,y \in \mathfrak{R}^n$ and $z \in \mathfrak{R}^{2n}$ is an additive information cost comparison if $\mathcal{G}^{add}(x,y) \equiv \mathcal{G}^{add}(x \oplus y) = \mathcal{G}^{add}(x) + \mathcal{G}^{add}(y)$.

There are a number of possibilities which the author discusses in [12]

- *Relative Entropy* or *Kullback-Leibler information* :

$$I(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log_2 \left(\frac{p_i}{q_i} \right) \quad (5.1)$$

where the following convention holds: $\log_2(0) = -\infty$, $\log_2(x/0) = +\infty$ for $x \geq 0$. This measure is asymmetrical which may be preferred in some applications. If symmetry is required, the *Kullback divergence (Relative Entropy)* should be used.

$$J(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log_2 \left(\frac{p_i}{q_i} \right) + \sum_i q_i \log_2 \left(\frac{q_i}{p_i} \right) \quad (5.2)$$

The more simple measure of

- l^2 distance

$$l(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2^2 \quad (5.3)$$

i.e. the square of the l^2 norm. Note that this measure can be extended for all l^p space ($p \geq 1$).

- *Hellinger distance*

$$h(\mathbf{p}, \mathbf{q}) = \sum_i (\sqrt{p(i)} - \sqrt{q(i)})^2 \quad (5.4)$$

The Hellinger distance and the first measure – *Relative* or *Cross Entropy* both belong to the Csiszar *f-divergence* class of dissimilarity measures (see [1] for a comprehensive review), which ultimately measure the distance between two probability distributions by considering the dispersion of their likelihood ratios.

These discriminant measures are used to decide which nodes out of all the possible subspaces carry the most useful information and it is inferred from this criterion which branches of the balanced binary tree are kept/pruned. Clearly these cost functions operate on the comparison of children and parent sub-nodes but what quantity composes these vectors? They are not simply the wavelet/cosine packet decompositions. Such a quantity was admissible in the previous chapter where expansion of signals in their own best bases was the aim. For the purposes of classification, a quantity that reflects the global characteristics of a given class is desirable. Possible quantities include the probability distribution function of a particular class which would be a useful criterion upon which to calculate the Discriminant Energy Concentration Measures (DECM). Another could invoke the use of Cumulative Distribution Functions (*cdfs*), as discussed in [4] where it is used due to its simplicity in computing empirically than the average Probability Distribution Function (*pdf*). The measure used in the original LDB is based on the mean of the class energies, called a Time-Frequency Energy (TFE) Map.

Definition 5.3. Let $\{x_i^{(y)}\}_{i=1}^{N_y}$ be a training set belonging to class y , N_y is total number of signals in class (y) training set. Then the Time Frequency Energy map of class y is a packet table of values indexed by the triplet (j, k, m).

$$TFE_{y(j,k,m)} = \frac{\sum_{i=1}^{N_y} (w_{j,k,m} x_i^{(y)})^2}{\sum_{i=1}^{N_y} \|x_i^{(y)}\|_2^2} \quad (5.5)$$

for $j = 0, \dots, J, k = 0, \dots, 2^j - 1, m = 0, \dots, 2^{n-j} - 1$

For a given class (y), TFE_y is computed by the accumulation of the squares of the expansion coefficients for all signals in a class at each (j, k, m) in the packet table followed by normalisation by the total energy in the class. It should be noted that this normalisation is particularly important for real-world applications like speech classification where typically one is faced with significantly different numbers of signals per class.

The cost measures, generically termed $\mathcal{G}(\cdot)$, are then combined amongst y sequences to result in an overall discrepancy for a given subspace. Definition 5.1, can be expanded to include any number of classes. e.g. for the two class case, definition (5.1) becomes $\mathcal{G}^{add}(\{x, y\}) = \sum_i \mathcal{G}^{add}(x_i, y_i)$. Using the additivity of the $\mathcal{G}(\cdot)$, equation (5.6) combines any number of

classes into a single measure

$$\mathcal{G}^{add}\left(\{x^{(y)}\}_{y=1}^{N_y}\right) = \sum_{m=1}^{N_y-1} \sum_{n=m+1}^{N_y} \mathcal{G}^{add}(x^m, x^n) \quad (5.6)$$

Here is the final LDB algorithm used in the experiment. Assuming $\mathcal{G}_{j,k}$ is the discriminant measure at a particular subnode, whether additive or not and let $D_{j,k}$ represent the Best Discriminant Basis and $R_{j,k}$ the fully expanded, redundant basis. Then

- (i) Choose to use either trigonometric dictionaries or Wavelet Packets for the transform.

- (ii) Expand every signal in the training set into its wavelet packet table.
- (iii) Determine the set of most discriminant subspaces using a top down pruning methodology by testing the efficacy of each subspace for discrimination.

the process is as follows

set a temporary array $\mathfrak{S}_{j,k} = \mathcal{G}_{j,k}$

if $\mathfrak{S}_{j,k} \geq \mathfrak{S}_{j+1,2k} + \mathfrak{S}_{j+1,2k+1}$; $D_{j,k} = R_{j,k}$; else

$D_{j,k} = D_{j+1,2k} \oplus D_{j+1,2k+1}$ and set $\mathfrak{S}_{j,k} = \mathfrak{S}_{j+1,2k} + \mathfrak{S}_{j+1,2k+1}$

- (iv) Rank the expansion coefficients according to their discriminant power and from these select the top $k \leq n$ features (where $n = 2^{no}$ is the dyadic length of the signal) for each signal in the training class to construct the final classifier.

The LDB gained from step two is an orthonormal basis, also if the cost function is additive, this step will be fast.

Step 4 is not crucial to the success of the algorithm since theoretically one can still design the classifier on all the features; however, if the dimensionality of the problem is reduced, this step will reduce the number of interfering components in the decomposition, making the class-specific features more robust. Computational training times will simultaneously be reduced.

In practice, one can rank the expansion coefficients using a number of approaches:

- a) Find the discriminant validity of a particular basis function in the LDB expansion, for example invoke the original cost function $\mathcal{G}(\cdot)$ on the wavelet coefficients expanded in the new basis.
- b) Use Fisher's class separability index, as described in Section (5.3.2) to rank the coefficients.

5.3.2 Non-Additive Criteria

As an extension to the above method, one can implement a non-additive cost whereby Step 3 of the LDB can be modified as follows:

Determine the set of most discriminant subspaces as before except:

if $\mathfrak{S}_{j,k} \geq \mathfrak{S}_{j+1,2k} \cup \mathfrak{S}_{j+1,2k+1}$; $D_{j,k}=R_{j,k}$; else

$D_{j,k}=D_{j+1,2k} \oplus D_{j+1,2k+1}$ and set $\mathfrak{S}_{j,k} = \mathfrak{S}_{j+1,2k} \cup \mathfrak{S}_{j+1,2k+1}$

i.e. the discriminant performance of a parent node with the *union* of its two children subnodes is instead considered. The selection scheme is thus rather different to the standard LDB with additive costs. In fact, if the cost function is the actual misclassification rate, then using the divide and conquer strategy described in Section 4.3, one is not guaranteed a basis selection from the $2^{2^{J-1}}$ possible bases that necessarily minimises the misclassification rate. This is because the classification error gained from the union of two subspaces that are individually best by themselves will not necessarily be smaller than the error of the union of two subspaces that are individually not best (see [9] for more information on this issue that arises generally in feature extraction based on misclassification errors).

Instead of misclassification rate, these experiments introduce a different approach - they use a non-additive proxy for misclassification related to the Mahalanobis distance between classes. To consider this and its relationship to overall LDA classification, LDA shall be briefly reviewed [8].

One can write the sample mean of signals belonging to class C_c as

$$\mathbf{m}_c = \frac{1}{N_c} \sum_{\mathbf{x} \in C_c} \mathbf{x} \quad (5.7)$$

The between class scatter S_B can be written as:

$$S_B = \sum_{c=1}^C \pi_c (\mathbf{m} - \mathbf{m}_c)(\mathbf{m} - \mathbf{m}_c)^t \quad (5.8)$$

π_c is the prior probability of signal membership to a class, m is the total mean for all signals in the training classes and m_c is the total mean of all training signals belonging to a class c . A suitable estimate for the prior probability is N_c / N . The idea of LDA is to maximise a separability statistic between the classes, however the consideration of S_B on its own is not a sufficient criterion. Maximising this may indeed give best separability between classes however overlap of the individual covariance matrices, the within class scatter - S_W of each class may still occur, see Figure 5.1.

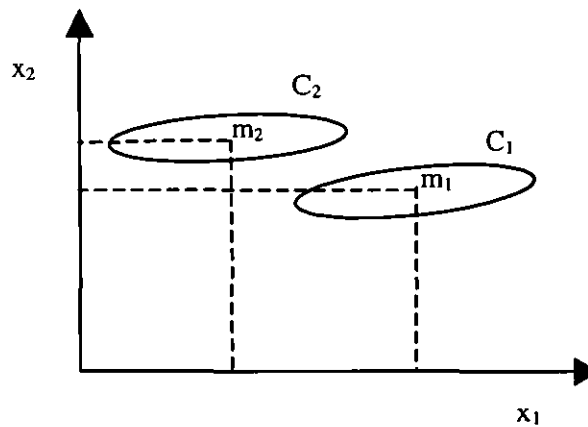


Figure 5.1 Showing importance of considering within class covariance or spread when maximising a separability criterion. The idea of LDA is to find a view of the data that best shows the difference between the training data classes. If S_B alone is used, then as indicated in the figure, the data would become projected onto the x_1 axis because this gives the maximum for between class separation. For the purposes of discrimination, however, projection onto the x_2 axis is most desirable. This is only possible when within class spread is jointly considered.

$$S_W = \frac{1}{N} \sum_{c=1}^C \sum_{i=1}^{N_c} (x_i^{(c)} - m_c)(x_i^{(c)} - m_c)^t \quad (5.9)$$

If $\tilde{X} = A'X$ denotes a linear transformation of the original training exponents, S_B and S_W can just be written $\tilde{S}_B = A'S_B A$ and $\tilde{S}_W = A'S_W A$. Fishers original linear discriminant which he developed in 1936 aimed to maximise the 'between to within' class variance ratio. This is achieved in practice by finding the ratio of the determinants of \tilde{S}_B and \tilde{S}_W :

$$J(A) = \frac{A' S_B A}{A' S_W A} \quad (5.10)$$

Which requires solving the following eigenvalue problem, the solution to which is found by forming the *canonical variates* of λ

$$S_B A = \lambda S_W A \quad (5.11)$$

In this chapter, the Mahalanobis distance used to weight the discriminant power of a subspace is closely related to LDA. In fact it can be directly used as an alternative yet closely related evaluation of between and within class distance in LDA. This approach is followed here to try and provide a seamless link between best basis choice – feature extraction for best discrimination and final classification rate.

The Mahalanobis distance $\Delta_{i,j}^2$ between two classes i and j is defined as:

$$\Delta_{i,j}^2 = (\mathbf{m}_i - \mathbf{m}_j) \Sigma^{-1} (\mathbf{m}_i - \mathbf{m}_j)^t \quad (5.12)$$

where Σ is called the pooled covariance matrix of all the training data:

$$\Sigma = \sum_{c=1}^C (N_c - 1) S_c / (N - C) \quad (5.13)$$

with S_c equal to the within class scatter for class c , N is the total number of signals across classes, C is the number of classes.

Naturally, the $\Delta_{i,j}^2$ are computed recursively on each subspace in the wavelet/cosine packet table. The final cost therefore, measures the goodness of training class separability of each subspace, and the final basis chosen will attempt to maximise this separation. This approach is related to that described in [15], where the Local Regression Bases for non-additive costs is developed. However, the final misclassification rate of an LDA classifier is used instead as a cost functional. The cost used here $\Delta_{i,j}^2$, may constitute a closer measure of operating

system characteristics, i.e. class separation in the linear sense. It is further noted in [15] that LRB using LDA is seen to be slightly worse; furthermore, training times are considerably increased since LRB involves evaluating LDA at each of the $(2^{(J+1)}-1)$ subspaces which requires expensive eigenvalue $O(n^3)$ calculations in addition to working out within and between class scatter. This results in a prohibitive training phase when dealing with large samples of real world signals as in the case of speech signals. Saito tests the efficacy of his algorithms using synthetic signals, or on seismic trace data – a somewhat different setting to the current work in which the number of training samples is typically low compared with signal dimensionality, which is generally high.

Employing an LDA related cost function prior to subspace evaluation yields an approach linking feature extractor to the final classifier. This philosophy could similarly be extended to include any type of classifier, indeed it is easily shown that there exists a close relationship extending Fisher's discriminant criterion to the least squares criterion commonly used in training Artificial Neural Networks (ANNs), [2]. One could certainly use a measure based on ANNs instead of the Mahalanobis distance e.g. similar to [10]. Let us now look more closely at the relationship between using Mahalanobis distance compared with some existing cost measures. In [17], such measures have been categorised into four broad classes, see also [16]:

- (i) As in the original LDB, a measure based on the distance between the mean class energies.
- (ii) A measure based on differences among the pdfs of each class (implemented in [17]).
- (iii) A measure based on differences amongst the cumulative distribution functions (cdf's) of each class [4].
- (iv) A measure, derived at each subspace, based on the actual misclassification rate of the entire training set projected in that vector, similar to that examined in this section.

Using (iv) is similar to LRB and is the nearest to using $\Delta_{i,j}^2$. However (iv) is equivalent to evaluating the specific quantitative goal – misclassification rate and is thus more suitable for picking out bases that give the best distinction amongst classes instead of, as in standard LDB a proxy for class separability (the relative entropy). Furthermore, in original LDB, a distance measure which tries to separate amongst the *energy* of the wavelet coefficients may

not be the most suitable for discrimination as pointed out in [17]. Instead, the authors here use a Type (ii) measure although in practice a good pdf estimate is more difficult to obtain. They go on to invoke additive cost functionals as in the original LDB. A related method using the Matching pursuit instead of Best Basis (called the Discriminant pursuit) was used in [3] where the goal is similar to ours – improvement of LDA using a 1-D form of Fishers criterion.

5.4 Results

In the following experiments, the first algorithm to be implemented was the standard Local Discriminant Basis algorithm (LDB), using an additive cost function of Relative Entropy. The best $k \leq n$ features were chosen using the same criterion. This approach was compared with a configuration related to the Local Regression Basis algorithm ((LRB), which is essentially LDB with non-additive costs) using the Mahalanobis distance measure as a cost function and the expansion coefficients ranked, and signal dimension reduced according to this class separability criterion. In addition, for MLRB, a small non-linear thresholding to the subspace vectors prior to calculating this distance was applied (about 0.8 of the total number of coefficients present in parent and child nodes). The final classifier in all cases was LDA. Thus a closely related optimality criterion was used both in the evaluation of suitable features for class separability as in the final classification estimate. In the study, both wavelet and cosine packets were considered. In the latter case two wavelet packets were chosen that were empirically determined from Chapter 3 to give the best performance, namely the Daubechies filter of 6 vanishing moments and the Coiflet filter with 10 vanishing moments depending on the problem, see Table 5.1.

The two phoneme classification problems analysed here are the well separated vowels /aa/, /ax/, /iy/ corresponding to the back, mid and front positions of the tongue during voicing were examined, although the other phoneme classification problems are included for completeness (Tables 5.2 and 5.3). The three unvoiced stops, /p/, /t/, /k/ were discriminated against one another to evaluate the performance of the two methods. In both cases, the phonemes were extracted from dialect regions 1 and 2 of the TIMIT database from all speakers both male and female to ensure a good statistical representation of each sound. The

speech datasets used were sampled at a rate of 16kHz; thus the 32ms window used, comprised of ~512 samples.

The results gained using the methods outlined plus a benchmark version of the STFT, are given in Table 5.1 through 5.3.

Technique		Error Rate (Training)	Error Rate (Testing)	Problem
LDA on STFT64		16.48%	22.38%	/iy/,/aa/,/ax/
LDA on LDBuRE 64	WP-D6	8.74%	8.81%	/iy/,/aa/,/ax/
	CP	8.62%	9.15%	
LDA on LRBuMD 64	WP-D6	8.51%	8.25%	/iy/,/aa/,/ax/
	CP	8.86%	9.02%	
LDA on STFT64		39.16%	46.13%	/p/,/t/,/k/
LDA on LDBuRE 64	WP-C10	32.15%	40.32%	/p/,/t/,/k/
	CP	28.69%	41.94%	
LDA on LRBuMD 64	WP-C10	25.45%	39.68%	/p/,/t/,/k/
	CP	28.48%	40.65%	

Table 5.1: Misclassification rates of the feature extraction techniques when applied to two phoneme classification problems. Quantities in bold show the best overall classification rate achieved which in both studies was on top 64 coefficients with standard LDB on wavelet packets. LDA, STFT64, LDB64uRE indicate respectively, the type of final classifier used, top 64 Short-Time Fourier Transform gained from the whole 512, the top 64 expansion coefficients extracted using standard LDB with Relative Entropy as the cost function. LRBuMD64 is the top 64 co-ordinates obtained using a non-additive Mahalanobis Distance measure. CP and WP indicate whether wavelet or cosine packets were used, where C10 stands for the Coiflet wavelet packet with 10 vanishing moments.

Technique		Error Rate (Training)	Error Rate (Testing)	Problem
LDA on STFT64		52.71%	54.76%	/m/,/n/,/g/
LDA on LDBuRE 64	WP-D8	40.50%	46.40%	/m/,/n/,/g/
LDA on LRBuMD 64	WP-D8	40.33%	45.53%	/m/,/n/,/g/
LDA on STFT64		32.82%	33.73%	/w/,/y/,/l/,/r/
LDA on LDBuRE 64	WP-S7	20.85%	24.93%	/w/,/y/,/l/,/r/
LDA on MLRBu MD64	WP S7	20.62%	25.82%	/w/,/y/,/l/,/r/

Table 5.2: Misclassification rates of the feature extraction techniques when applied to the two phoneme classification problems of nasal stops and semivowels. Quantities in bold show the best overall classification rate achieved which in both studies was on top 64 coefficients with standard LDB on wavelet packets. LDA, STFT64, LDB64uRE indicate respectively, the type of final classifier used, top 64 Short-Time Fourier Transform gained from the whole 512, the top 64 expansion coefficients extracted using standard LDB with Relative Entropy as the cost function. LRBuMD64 is the top 64 co-ordinates obtained using a non-additive Mahalanobis Distance measure. WP indicates that wavelet packets were used in the experiment, where D8 stands for the eighth order Daubechies wavelet packet and S7 for a seventh order Symmlet packet.

Technique		Error Rate (Training)	Error Rate (Testing)	Problem
LDA on STFT64		23.53%	23.54%	/f/,/T/,/s/
LDA on LDBuRE 64	WP-D7,J=5	17.29%	17.64%	/f/,/T/,/s/
LDA on MLRBu MD 64	CP,J=5	17.48%	18.08%	/f/,/T/,/s/
LDA on STFT64		23.84%	26.18%	/v/,/dh/,/z/
LDA on LDBuRE 64	WP-C10	15.83%	17.49%	/v/,/dh/,/z/
LDA on MLRBu MD64	WP C10	15.83%	16.82%	/v/,/dh/,/z/

Table 5.3: Misclassification rates of the feature extraction techniques when applied to the two phoneme classification problems of unvoiced and voiced stops. Quantities in bold show the best overall classification rate achieved which in both studies was on top 64 coefficients with standard LDB on wavelet packets. LDA, STFT64, LDB64uRE indicate respectively, the type of final classifier used, top 64 Short-Time Fourier Transform gained from the whole 512, the top 64 expansion coefficients extracted using standard LDB with Relative Entropy as the cost function. LRBuMD64 is the top 64 co-ordinates obtained using a non-additive Mahalanobis Distance measure. WP indicates where wavelet packets were used in the experiment, D7 stands for a Daubechies order 7 wavelet and C10 stands for the Coiflet wavelet packet with 10 vanishing moments. J=5 indicates the level of decomposition for the /f/,/T/,/s/ problem

5.5 Discussion

The discussion will start by noting that the 'feature selection' rule used in both LDB and modified LRB was the same as the discriminant measure which determined the best-subspace itself in both cases. With regard to the best $k < n$, a commonly used rule of thumb of approximately 10% of the original signal dimensionality was chosen.

The performance of the modified LRB algorithm was seen to outperform the original LDB algorithm for both cases examined (with this particular choice of k) and also in both cases, wavelet packets do best. The reader may also note that both the original LDB and modified LRB give results improving on those given for the /iy/, /aa/, /ax/ problem of Chapter 4; modified LRB in this case does as expected, namely it has successfully extracted a basis to maximise the relation of (5.12). This may be seen by examining training performance, which is a measure of how well the basis extracted linearly separates the signal. The result of 8.51% for modified LRB using wavelet packets is the best of all. The prediction performance of 8.25% is also an improvement over standard LDB of around 9.8% reduction in misclassification rate. The fact that the misclassification rate is relatively low indicates that this problem is linearly separable. It therefore makes sense to apply modified LRB in this case (closeness of training and testing results indicates that good generalisation has been achieved in this method).

In the case of the /p/, /t/, /k/ sounds, however, prediction performance is slightly degraded compared to Chapter 4: Testing misclassification rates of 38.06% with StdBB using entropy compared with 39.68% with modified LRB using Mahalanobis distance although modified LRB still outperforms LDB contrary to the results reported for the original LRB in [15]. However, a significant improvement is seen in this case when the training rates are considered: 32.15% for Best-Basis compared with 25.45% for modified LRB using Mahalanobis distance. Normally, it is desirable to have the difference between training and testing rates as small as possible since this signifies a classifier to have learnt well the features of a particular problem. One would also normally associate, for a difficult problem like this where generalisation is poor, a jump in training performance as indicative of over-

5.5 Discussion

The discussion will start by noting that the ‘feature selection’ rule used in both LDB and modified LRB was the same as the discriminant measure which determined the best-subspace itself in both cases. With regard to the best $k < n$, a commonly used rule of thumb of approximately 10% of the original signal dimensionality was chosen.

The performance of the modified LRB algorithm was seen to outperform the original LDB algorithm for both cases examined (with this particular choice of k) and also in both cases, wavelet packets do best. The reader may also note that both the original LDB and modified LRB give results improving on those given for the /iy/, /aa/, /ax/ problem of Chapter 4; modified LRB in this case does as expected, namely it has successfully extracted a basis to maximise the relation of (5.12). This may be seen by examining training performance, which is a measure of how well the basis extracted linearly separates the signal. The result of 8.51% for modified LRB using wavelet packets is the best of all. The prediction performance of 8.25% is also an improvement over standard LDB of around 9.8% reduction in misclassification rate. The fact that the misclassification rate is relatively low indicates that this problem is linearly separable. It therefore makes sense to apply modified LRB in this case (closeness of training and testing results indicates that good generalisation has been achieved in this method).

In the case of the /p/, /t/, /k/ sounds, however, prediction performance is slightly degraded compared to Chapter 4: Testing misclassification rates of 38.06% with StdBB using entropy compared with 39.68% with modified LRB using Mahalanobis distance although modified LRB still outperforms LDB contrary to the results reported for the original LRB in [15]. However, a significant improvement is seen in this case when the training rates are considered: 32.15% for Best-Basis compared with 25.45% for modified LRB using Mahalanobis distance. Normally, it is desirable to have the difference between training and testing rates as small as possible since this signifies a classifier to have learnt well the features of a particular problem. One would also normally associate, for a difficult problem like this where generalisation is poor, a jump in training performance as indicative of over-

adaptation and consequently, it would be expected that the prediction performance would fall correspondingly. However this has not occurred, indicating perhaps that the linear separability of the training features has certainly been much improved although generalisation is still difficult but without suffering an appreciable fall in prediction rates.

Figure 5.2 compares the covariance matrices of the training data derived via LDA before and after feature extraction for the /iy/, /aa/, /ax/ problem. Before feature extraction and dimensionality reduction, they show little structure beyond noise except at low frequencies and are difficult to interpret. After implementing modified LRB using Mahalanobis costs, the variates are much less noisy, showing more low frequency structure. Most of the signal energy has been compacted into a top few co-ordinates on the diagonal – indicating these methods have extracted independent features suitable for classification. It is worth comparing original LDB for the same variates. Modified LRB can be seen as having improved diagonalisation of the covariance matrix (Figure 5.3).

Figures 5.4 and 5.5 show the comparative performance of the two methods as the value of k is varied. One can note in the case of Figure 5.4 that LDB does significantly better on a small number of features compared to MLRB whose performance is best when k is around 10% of the total signal dimension. The training features for both classification problems are more independent when extracted via MLRB however, Figure 5.5 in particular shows that this is at the expense of worse generalisation capability compared with LDB. This could perhaps be overcome by gaining the subspace cost for MLRB using a validation data-set rather than the training data-set.

The initial computational cost of LRB related methods are significantly greater than LDB due to the reasons mentioned in Section 5.1.2. However this is only a training cost and once a basis tree is worked out, all subsequent signal known to belong to that phonetic subclass can be decomposed in a comparably fast manner to standard Best Basis procedures.

The type of system proposed has been shown to provide some improvement over standard LDB in certain situations. As a pre-processing technique to other standard classifiers e.g. (Hidden Markov Models) HMM's or ANN's it shows promise.

5.6 Summary

This chapter introduced a new discriminant feature extractor which captures local signal attributes to enhance the performance of the final classifier. The algorithm which was called the Modified Local Regression Basis algorithm (MLRB) compared well with a related method which uses an additive cost function - the Local Discriminant Basis algorithm (LDB) and was seen to provide features that helped to diagonalise the covariance matrix of the training data. This was shown to be equivalent to obtaining more independent features and thus training misclassification performance for MLRB was appropriately decreased without much loss of generality by the classifier. Both discriminant techniques were also seen to give similar performance to the techniques examined in Chapters 3 and 4.

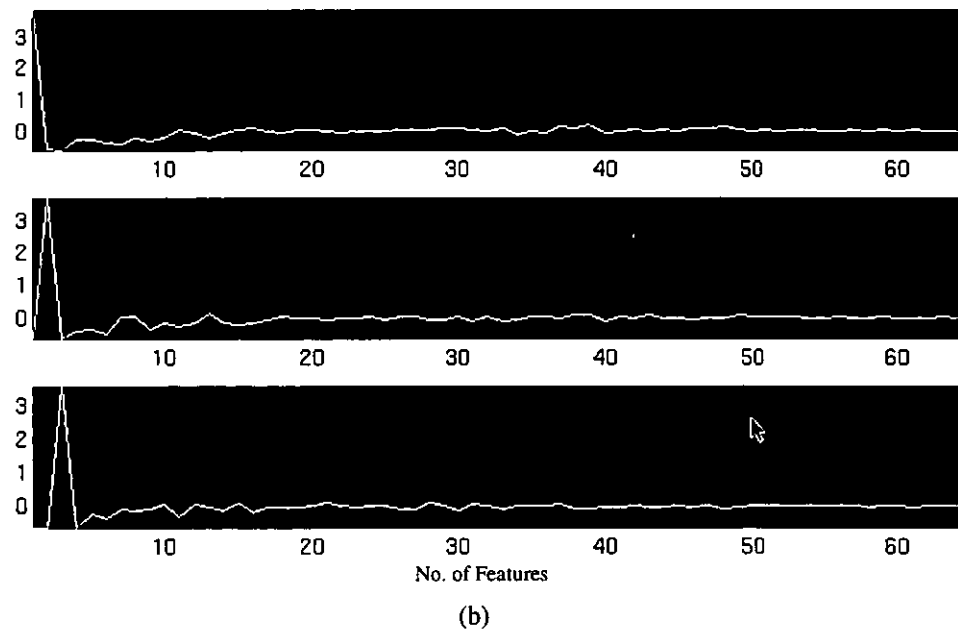
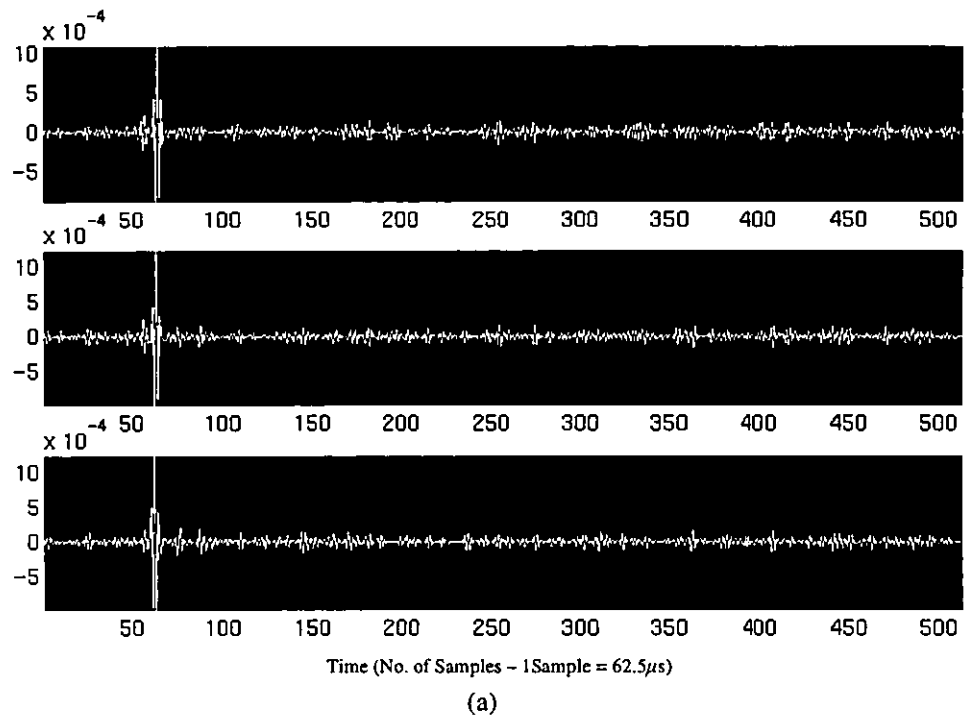
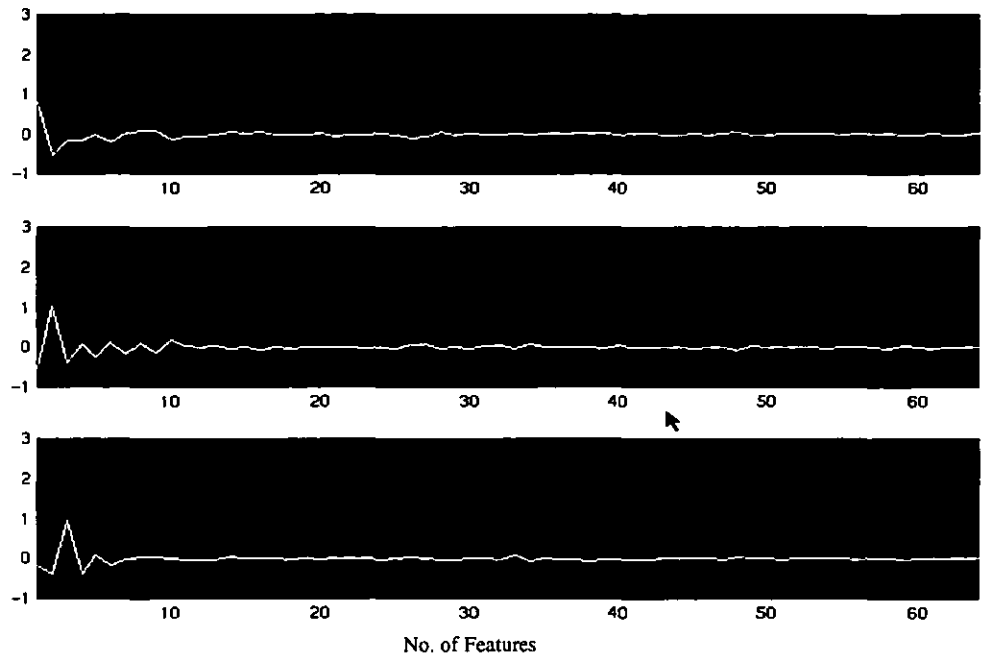
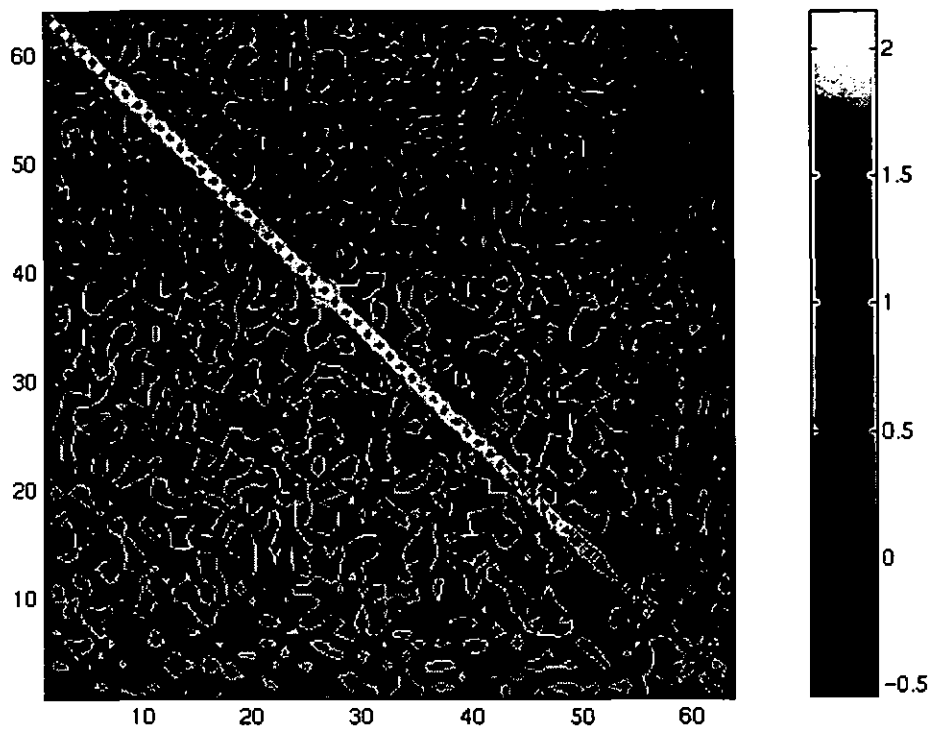


Figure 5.2 –See Overleaf for details

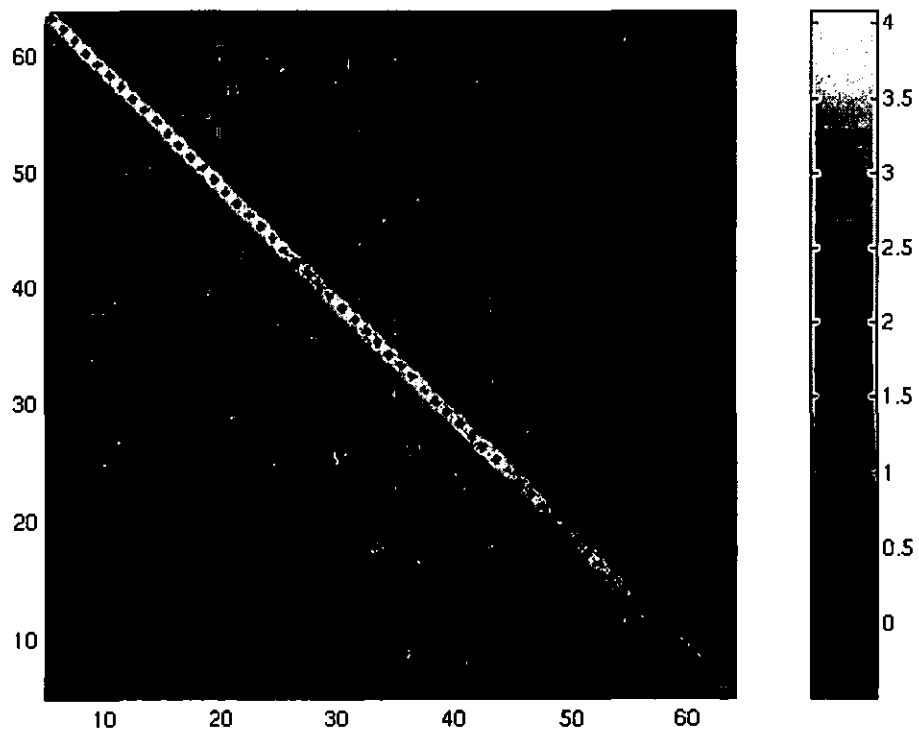


(c)

Figure 5.2 First three rows of covariance matrix Σ for the /iy/,/aa/,/ax/ problem from (a) Original Euclidean bases (original signal) (b) Top 64 co-ordinates derived via the Daubechies wavelet with 6 vanishing moments and the modified LRB algorithm using the Mahalanobis distance measure and (c) The original LDB algorithm with a Daubechies 6 wavelet and Relative Entropy as the discriminant energy concentration measure. Note the noisiness of (a) compared with (b) and (c) where the discriminant energies are well packed into the top few coefficients. Note also the modified LRB algorithm in which significantly more energy is present in the first few co-ordinates. This, and the fact the matrix is better diagonalised results in superior overall training performance for this method although prediction performance is correspondingly decreased.



(a)



(b)

Figure 5.3 (a) Pooled covariance matrix of the training data using the top $k=64$ most discriminant expansion coefficients.(b) Corresponding modified LRB matrix showing significantly higher concentration of energy along the diagonal. When this matrix is diagonalised, eqn.(5.12) reduces to the square of the Euclidean distance.

Problem is $/p/,t/,/k/$.

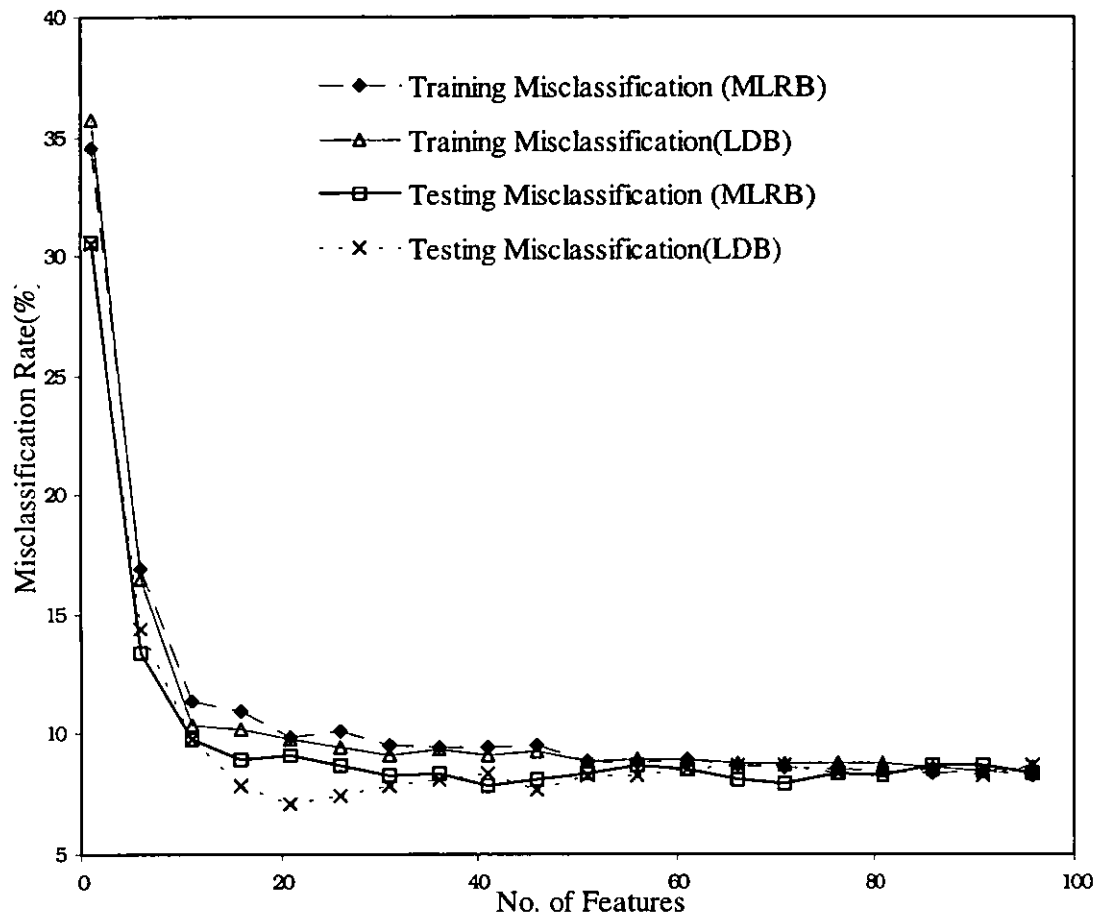


Figure 5.4 Misclassification rates of /iy/,/aa/,/ax/ problem vs. number of discriminant features supplied to the classifier using LDB and MLRB respectively. MLRB outperforms LDB when around 10% of features are used, however LDB does significantly better when only ~4% of features are used.

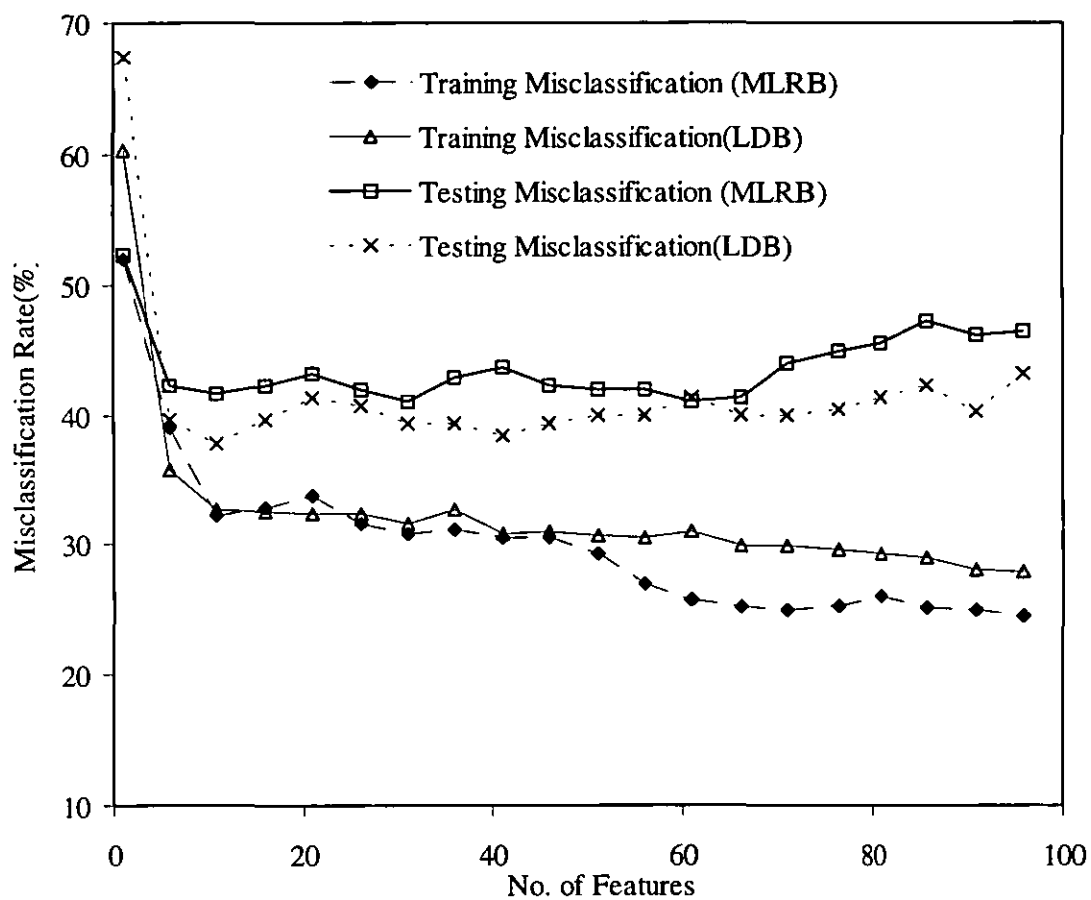


Figure 5.5 Misclassification rates of /p/, /t/, /k/ problem vs. number of discriminant features supplied to the classifier using LDB and MLRB respectively. One can see higher adaptation to the training data using MLRB. This affects the generalisation of this method compared to LDB indicating a certain amount of over adaptation has occurred.

5.7 References

- [1] Basseville, M., "Distance Measures for Signal Processing and Pattern Recognition," in *Signal Processing* 18 (1989) 349-369.
- [2] Bishop, C.M., "Neural Networks for Pattern for Pattern Recognition." *Oxford University Press*. (1997)
- [3] Buckheit, J. and Donoho, D.L., "Improved Linear Discrimination Using Time-Frequency," *Technical Report*, Stanford University, 1995.
- [4] Buckheit, J. . and Donoho, D.L., "Time-frequency Tilings which best expose the Non-Gaussian behaviour for a Stochastic Process," *Proc. International Symposium on Time-Frequency and Time-Scale Analysis, IEEE*, June 18-21, 1996, Paris, France.
- [5] Coifman, R.R. and Wickerhauser, M.V. "Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
- [6] D'Alessandro, C. and Richard, G. "Random wavelet representation of unvoiced speech," *Proc. IEEE-SP Int.Symp. on Time Frequency and Time Scale Analysis*, pp. 41-44 (Oct.1992).
- [7] Daubechies, I. "Orthonormal bases of compactly supported wavelets," *Comm in Pure and Applied Math.*, vol.41 No.7, pp.909-996, 1988.
- [8] Duda, R.O. and Hart, P.E., "Pattern Classification and Scene Analysis." *New York, John Wiley*.(1973)
- [9] Fukunaga, K., "Introduction to Statistical Pattern Recognition." *Academic Press, San Diego, CA*. 1990.Second Edition.

- [10] Huynh, Q., Cooper, L.N., Intrator, N., Shouval, H., "Classification of Underwater Mammals using Feature Extraction Based on Time-Frequency Analysis and BCM Theory." *submitted to IEEE Trans. On Signal Processing*
- [11] Kadambe, S. and Faye Boudreaux-Bartels, G. "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
- [12] Long, C.J. and Datta, S., "Time-Frequency Dictionaries for Improved Local Feature Extraction," *Proc. of IEE Colloquium on Pattern Recognition*, London, 26 Feb. 1997, pp 9/1 – 9/6.
- [13] Mallat, S. "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 674-693, 1989.
- [14] Rioul, O. and Vetterli, M. "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October 1991.
- [15] Saito, N. "Local Feature Extraction and its Applications using a Library of Bases," *A Dissertation*, Yale University, Dec. 1994.
- [16] Saito, N., "Classification of Geophysical Acoustic Waveforms using Time-Frequency Atoms," *Proceedings of Statistical Computing, Amer. Statist. Assoc.* (1996).
- [17] Saito, N., "Improved Local Discriminant Bases Using Empirical Probability Density Estimation." *Proceedings of Statistical Computing, Amer. Statist. Assoc.* (1996).
- [18] Thiripuraneni, J.A. et al, "Mixed Malvar Wavelets for Non-stationary Signal Representation," *Proc. ICASSP vol. 1 pp. 13-16, Atlanta* (1996).
- [19] Wickerhauser, M.V., "Fast Approximate Factor Analysis," *Curves and Surfaces in Computer Vision and Graphics II*, Oct. 1991, Proc. SPIE 1610, pp. 23-32.
- [20] Wickerhauser, M.V. "Acoustic Signal Compression with Wavelet Packets," In Chui C.K.(ed) *Wavelets: A Tutorial in Theory and applications* (1992).

CHAPTER SIX

Conclusions

6.1 Overview

This thesis has explored different ways of evaluating recognition features in speech signals using wavelets and their relatives. There were three main experimental chapters each considering a different approach to utilising the properties of wavelets in the setting of phoneme classification. The evaluations carried out demonstrated the importance of understanding the mathematical properties of wavelets. Wavelets now compose a field of mathematics that is now quite mature and well-understood where it is widely recognised that in certain situations it makes better sense to use wavelets rather than conventional techniques. This could be due to a number of reasons. For example in compression of real world signals and images, wavelets provide well-localised time-frequency information, and can take advantage of the closely correlated internal structures that exist in real-world signals resulting in a more compact representation. Wavelets are therefore good descriptors in their ability to approximate, but when it comes to applications such as

feature extraction for speech classification where and why might they be seen to provide improvement?

Chapter 3 saw the straightforward replacement of Fourier methods in the pre-classification stage of a speech recognition system by the standard dyadic wavelet transform. Overall, when the top ~10% of signal features were chosen by ranking the squares of their respective expansion coefficients wavelets, they did significantly better than the STFT with some basis functions more able to provide good approximation than others. This was found to be explicable when one considered the Lipschitz regularity of the two methods from which it was subsequently recognised that the rate of decay of the wavelet transform was faster in the neighbourhood of high frequency events than Fourier techniques. This phenomena is related to the zooming property of wavelets which allows multiresolutional analysis of a signal.

Chapter 4 considered the application of the Best-Basis algorithm originally designed for compression to the same set of problems. The Best Basis algorithm or more generally, the Best Basis paradigm extends the desirable attributes of wavelets to provide a framework which adapts according to the information content of the signal. It was noted that this property corresponded to an adaptive tiling of the time frequency plane where the Heisenberg windows were well adjusted to the spectra they contain.

Chapter 5 saw the development of the Best Basis paradigm for classification. In this setting, it is known as the Local Discriminant Basis algorithm, and here it was applied to the problem of speech classification. This approach differs fundamentally from the techniques described in Chapters 3 and 4 in that characteristics of the signal classes are taken into account when calculating the Best Basis. It was noted that distance measures such as Relative Entropy or Hellinger distance may not be an optimal choice for this particular problem. This framework was modified by viewing classification and feature extraction as one process, in that the classification methods itself decides which features are supplied to it by modelling class densities at each subspace and calculating class distances based on the measure used in the final classifier. This new approach was seen to

improve overall operating system characteristics (misclassification rate) under certain conditions.

The findings of this thesis can be summarised as follows.

6.1.1 Dyadic Wavelet Transform

- Reduces the misclassification rate on all phoneme classification problems examined – compared to Fourier techniques.
- The type of wavelet basis used affects the performance. Overall it was found that wavelets with a higher order of vanishing moments do better, due to the irregular nature of the speech signals.
- The features derived via wavelets retain spectral information such as formant frequencies. Distances derived from these features via LDA were seen to be directly related to differing points of articulation within the vocal tract, in the case of vowel sounds.
- The DWT provided best improvement on those groups of speech signals which contained transitory components. This was related to the fact that wavelets can differentiate between global and pointwise signal regularity (Lipschitz).
- Non-linear operations on wavelet features provide improvement over linear methods for speech classification problems.
- Dyadic wavelet bases offer a convenient framework for decomposing signals according to their time-frequency content, however the localisation characteristics are fixed as follows: high frequency events have narrow time support with broad frequency support. For low frequency events the converse is true. The results in this chapter motivated investigation into adaptive time-frequency tilings as described in Chapter 4.

6.1.2 The Best Basis Algorithm

- Overall improvement in misclassification rates compared with the DWT.
- In addition to the type of wavelet basis used, an extra parameter needs to be chosen *a priori*:- the cost function. The most robust overall choice for the speech problems was seen to be an l^1 cost measure. This is a typical choice for improving robustness in real world datasets which typically contain outliers.
- Additional robustness is related to translation invariance. The cycle-spinning algorithm was applied to our datasets to try and reduce this. Effectively, cycle spinning increases the numbers of samples in the training and testing datasets by producing circularly shifted versions of each signal. It provided some improvement in one of the two cases tried, indicating that some of the features used for recognition are, to some extent dependent on position. It was also noted that the number of spins required, or block size depended on the problem.
- It was noted that for transitory types of signal, the BB algorithm outperforms the DWT quite significantly. Two classification problems were chosen for comparison, one which was approximately smooth - /iy/,/aa/, and /ax/ and one which was more transitory in nature - /p/,/t/,/k/. The BB approach improved prediction performance for the /p/,/t/,/k/ problem and slightly worsened the /iy/,/aa/, and /ax/ case. Overall, in 5 out of the 6 classification examples tried, whether BB provided improvement or not, it was seen to have reduced the difference between training and testing performance. This indicated that the extra flexibility of the BB features were more easily learnt by the classifier.

6.1.3 Discriminant Wavelet Methods

Some of the variability within Best Basis searches mentioned in Chapter 4 is likely to have a bigger effect on real world problems like those investigated here, which contain large numbers of samples with random noise. This motivated an investigation into discriminant wavelet packet approaches where overall class structures could be modelled

and searched using wavelets and the Best Basis type searches. The LDB algorithm of Saito and Coifman was implemented and tested on the speech examples. The following was concluded:

- LDB did better than standard BB approaches in some cases and slightly worse in others although results were comparable. It was decided that this could be due to two factors. First, LDB uses the so-called time-frequency energy map to represent classes. This may be problematic in certain situations. For example, if one were attempting to discriminate between two phase shifted sine waves, the TFE (Time-Frequency Energy) or mean of both will be zero. Second, the discriminant measure of a sub-node is taken as an *approximation* of class contrast (using Relative Entropy or some such similar measure) when what one really needs is a more direct class separability measure.
- With this in mind, a new approach was developed where the Mahalanobis 'Between to Within' class scatter was measured and used as a cost. It was recognised that this cost was non-additive and as such was similar to the LRB (Local Regression Basis) algorithm of Saito although the difference is that his algorithm uses the final regression (misclassification) training error whereas our approach still concentrates on distances amongst classes.
- Modified LRB was seen to improve performance over LDB in 4 out of the 6 cases examined. This could be seen on examination of the pooled covariance matrix of the training data using modified LRB features; more of the discriminant training data energy was concentrated along the diagonal in fewer coefficients compared to LDB. However, this was only the case for $k \sim 10\%$ of signal length n (the benchmark number of features commonly used in discrimination problems). When k was varied, it was seen that although MLRB did better on the training data, it had over-adapted to the problem. It was noted that gaining the MLRB cost from a validation data-set instead of the training data itself could provide a solution to this.
- Modified LRB provided a closer relationship between feature extraction and classification stages compared to LDB.

6.2 Future Work

Wavelet based feature extraction makes theoretical sense to implement in speech recognition systems. Which type of approach to choose really depends on the problem. This thesis demonstrated that wavelets can prove useful if standard parameterisation techniques are simply substituted for corresponding wavelet approaches. Furthermore, with careful development, the wavelet transform has been extended to become yet more adaptive, and using the Library of Bases (Best Basis) paradigm offers an even more flexible framework. To fully take advantage of this, one now has the possibility of using techniques like LDB and modified LRB to gain a better understanding of the nature of the phonemes. In other words, these kind of techniques can be thought of as being able to iteratively peel off irrelevant information for discrimination (in a multiresolutional way), hence simplifying the problem. This should yield a different view of speech via these new features and we should try to interpret them in the same way as (Linear Predictive Coding) LPC and FFT parameters. The Best Basis paradigm is thus a useful research tool for speech so how else could one improve its performance? Forging closer links between discriminant wavelet approaches and the final classifier is another option; the above techniques should be extended to encompass more sophisticated classifiers such as Artificial Neural Networks (ANNs) or Hidden Markov Models (HMMs). Furthermore, the scope of the wavelet techniques should still provide comparable improvement if fundamental block sizes are increased, say, from phoneme to word. Thus further improvement of the modified LRB and LDB features could include work toward obtaining synchronous features suitable for HMM's or Recurrent Neural Networks.

APPENDIX

PUBLISHED PAPERS

Wavelet Based Feature Extraction for Phoneme Recognition

C.J.Long and S.Datta

Department of Electronic and Electrical Engineering

Loughborough University of Technology

Loughborough LE11 3TU, UK.

Email c.j.long@Lboro.ac.uk

ABSTRACT

In an effort to provide a more efficient representation of the acoustical speech signal in the pre-classification stage of a speech recognition system, we consider the application of the Best-Basis Algorithm of Coifman and Wickerhauser. This combines the advantages of using a smooth, compactly-supported wavelet basis with an adaptive time-scale analysis dependent on the problem at hand.

We start by briefly reviewing areas within speech recognition where the Wavelet Transform has been applied with some success. Examples include pitch detection, formant tracking, phoneme classification. Finally, our wavelet based feature extraction system is described and its performance on a simple phonetic classification problem given.

1. INTRODUCTION

Speech recognition systems generally carry out some kind of classification/recognition based upon speech features which are usually obtained via time-frequency representations such as Short Time Fourier Transforms (STFTs) or Linear Predictive Coding (LPC) techniques. In some respects, these methods may not be suitable for representing speech; they assume signal stationarity within a given time frame and may therefore lack the ability to analyse localised events accurately. Furthermore, the LPC approach assumes a particular linear (all-pole) model of speech production which strictly speaking is not the case.

Other approaches based on Cohens general class of time-frequency distributions such as the Cone-Kernel and Choi-Williams methods have also found use in speech recognition applications but have the drawback of introducing unwanted cross-terms into the representation.

The Wavelet Transform overcomes some of these limitations; it can provide a constant-Q analysis of a given signal by projection onto a set of basis functions that are scale variant with frequency. Each wavelet is a shifted scaled version of an original or mother wavelet. These families are usually orthogonal to one another, important since this yields computational efficiency and ease of numerical implementation. Other factors influencing the choice of Wavelet Transforms over conventional methods include their ability to capture localised features. Also, developments aimed at generalisation such as the Best-Basis Paradigm of Coifman and Wickerhauser [1] make for more flexible and useful representations.

We consider the possibility of providing a unified wavelet-based feature extraction tool, one designed to contend optimally with the acoustical characteristics particular to speech, in the most computationally efficient manner.

The indications are that the Wavelet Transform and its variants are useful in speech recognition due to their good feature localisation

but furthermore because more accurate (non-linear) speech production models can be assumed [2]. The adaptive nature of some existing techniques results in a reduction of error due to inter/intra speaker variation.

We shall begin by defining the wavelet transform.

2. WAVELETS AND SPEECH

2.1 The Discrete Wavelet Transform

The basic wavelet function $\psi(t)$ can be written

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

The Continuous Wavelet Transform is then defined as

$$X(\tau,a) = \frac{1}{\sqrt{a}} \int x(t) \psi\left(\frac{t-\tau}{a}\right) dt \quad (2)$$

where $\psi(t)$ is known as the *analysing* wavelet or *prototype* function. Typically, these continuous wavelet functions are overcomplete and therefore do not form a true orthonormal basis. Redundancy may be eliminated by appropriately sampling the wavelet on a dyadic lattice, i.e. in a manner that reflects the tiling of the time-frequency plane as in figure 1. An orthonormal basis of compactly supported wavelets can then be obtained to span $L^2(\mathfrak{R})$ (the space of all finite energy signals) by shifting and dilating the wavelet function $\psi(t)$ i.e.

$$\psi_{n,m}(k) = 2^{-n/2} \psi(2^{-n/2} k - mb_0) \quad (3)$$

where $n=1,2,\dots$ represents the scale and $m=0,1,\dots$ the time shift. Note that the scaling factor a is here chosen as 2 in order that the frequency axis is decomposed in octaves. Now if one chooses a suitable wavelet, a true orthonormal basis will be obtained. This results in a multiresolutional analysis of a given signal over $L^2(\mathfrak{R})$, yielding a time-scale decomposition similar to that exhibited in Figure 1. For further details on MRA, the reader is referred to the work of Mallat [3].

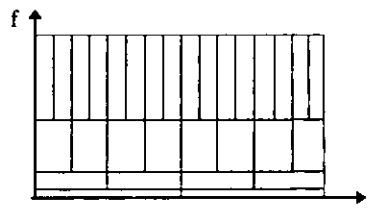


Figure 1: Tiling of time-frequency plane via the wavelet Transform.

2.2 Pitch and Formant Extraction using Wavelet Analysis

Kadambe & Boudreaux-Bartels [4] have used the multiresolutional properties of wavelets to propose an event-based pitch-detection system. Their method works by detecting the Glottal Closure Instant (GCI) and determines the pitch for each sample within a particular speech segment. This approach is particularly suitable for noisy speech.

Evangelista [5] has developed a 'Pitch-Synchronous' wavelet representation using a modified version of the QMF (Quadrature Mirror Filter) multiplexed filter bank outlined in [6]. Using the MRA properties of wavelets, the pitch-synchronous wavelet transform (PSWT) can be used for pitch tracking once the pitch has been extracted using conventional methods. Unique characterisation of speech events such as fricatives and occlusive unvoiced consonants may thus be achieved via the variation in the pitch of the signal.

Maes [7] reports success in the extraction of pitch and formants from speech. The speech signal is first decomposed into its subbands using the wavelet transform and the temporal behaviour of the speech in each subband is monitored using a 'squeezing' algorithm. Those components exhibiting similar temporal behaviour are then recombined and the resulting principle components represent the pitch and formant characteristics of the speech signal.

In [11], Wesfried introduces a speech representation based on the Adapted Local Trigonometric Transform. The window size into which the data is partitioned is dependent upon the spectrum it contains, and the transitions between windows is seen to be suitable for segmentation into voiced-unvoiced portions. A formant representation is also introduced by carrying out the following compression: locating and retaining the centres of mass for the highest-value peaks of the transform. From this, the local spectrum is said to represent the formant of the speech signal.

2.3 Phoneme and Speaker Classification using Adaptive Wavelets

The adaptive wavelet transform and the concept of the *super wavelet* were developed as an alternative to existing wavelet representation schemes [8]. Given a wavelet function of the form shown in (2), the idea is to iteratively find the translation and dilation parameters, τ and a respectively such that some application-dependent energy function is minimised. With respect to the classification problem, a set of wavelet coefficients would normally be estimated to represent certain features of a given signal. Classification can then be performed by using the feature set as the input to a neural net classifier. The adaptive wavelet based classifier is given as

$$v(n) = \sigma(u_n) = \sigma \left[\sum_{k=1}^K w_k \sum_{i=1}^T x_n(t) \psi \left(\frac{t - \tau_k}{a_k} \right) \right] \quad (4)$$

where $v(n)$ is the output for the n^{th} training vector $x_n(t)$ and $\sigma(z) = 1/[1 + \exp(-z)]$. For two classes, w_k, a_k, τ_k can be optimised by minimising the energy function in the least squares sense (see eq 5). In [2] then, two classification examples are considered with application to speech; classification of unvoiced phonemes and speaker identification.

$$E = \frac{1}{2} \sum_{n=1}^N (d_n - v_n)^2 \quad (5)$$

The system first models the phonemes using a mother wavelet similar to Figure 2 (used because of its noise-like characteristics) only of order 3 and then presents the wavelet features to a 2 layer feed-forward neural network. Speaker i.d. is similarly achieved only using a Morlet wavelet to model the phonemes since these are voiced and hence semi-periodic and smooth. The classifier then attempts to identify a speaker by clustering the associated utterances into one class. Results reported are very high accuracy, although exhaustive testing on a larger database will be needed to evaluate the method more accurately.

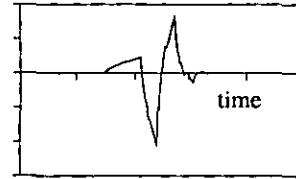


Figure 2: Daubechies Wavelet of order 4. This type of wavelet is used in Kadambe's unvoiced sounds classifier because of its suitability for modelling high frequency noise-like signals.

3. THE BEST-BASIS ALGORITHM

A generalisation of the Wavelet Transform originally designed for signal compression is the Best-basis algorithm first described in [1]. The idea is to do transform coding on a signal by choosing a wavelet basis which is best suited for the given problem, resulting in an adaptive time-scale analysis. In particular, two possibilities are proposed, the smooth local trigonometric transforms which essentially performs local Fourier analysis on the signal, and its frequency domain conjugate, the wavelet packet which similarly partitions the frequency axis smoothly. Since these transforms operate on recursively partitioned intervals on the respective axis, the bases whether wavelet packet or local trigonometric are said to form a *library of orthonormal bases*. If these bases are ordered by refinement, they form a tree which can be efficiently searched to result in only those coefficients which contain the most information.

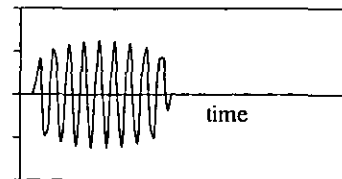


Figure 3: An example of a modulated smooth trigonometric packet. A localised sine dictionary, for example, would consist of a number of scaled, oscillatory versions of these.

In summary, the aim is to extract the maximum information or *features* from our signal by projection onto a co-ordinate system or basis function in which that signal is best (most efficiently) represented. What is meant by efficiency really depends on the final object. If compression is required, then the most efficient basis will be the one wherein most of the information is contained in just a few coefficients. On the other hand if we are interested in classification, a basis which most uniquely represents a given class

of signal in the presence of other known classes will be most desirable.

Figure 4 shows the structure of the wavelet based acoustic-phonetic feature extractor used in the pre-classification stage. Our library of basis contained just two dictionaries, wavelet packets and smooth localised cosine packets, although others are certainly possible. Thus the first stage of the system is to choose the most suitable of these for the problem at hand. This is done in practice by simply picking the one which gives minimum entropy among them [10].

3.1 Experimental

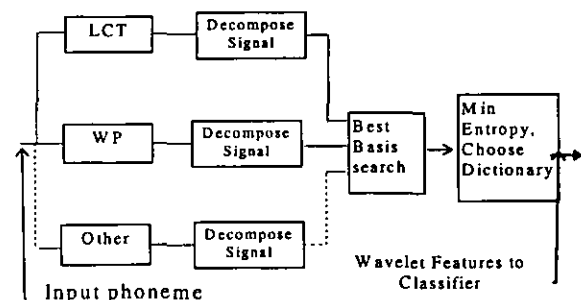


Figure 4: Wavelet-based feature extractor.

After Kadambe et al [2] who implement phonetic modelling and classification using adaptive wavelets, we use for training the voiced phoneme /d/ as in *had*, and the two unvoiced phonemes /s/ as in *ask* and /z/ as in *was*. These phonemes were extracted from a single male speaker in the TIMIT database. Each signal was low-pass filtered to resemble 8KHz band-limited telephone speech. The training features for the two layer feed-forward neural network were then obtained via the best-basis paradigm. A dictionary was chosen from our library of possible bases for each phoneme, dependent on which provided the minimum of a specified cost function, in this case entropy. As it turned out, the LCT (Local Cosine Transform) dictionary was selected for the voiced phoneme /d/ since these set of functions are smooth and most suitable for representing oscillatory signals. The /s/ and /z/ phonemes which correspond to different types of noise were best represented in terms of the wavelet packet with basis functions similar to Figure 2, i.e. a Daubechies wavelet of order 4. A fast search, (i.e. $O(n[\log n]^p)$ where $p=0,1,2$ depending on the basis type) was then performed in a binary tree similar to that of Figure 5. The wavelet features of the training vectors obtained using this method are shown in Figure 6(a), (b), and (c) along with the original signals decimated to a length of 1024 samples. A restriction, in fact, of this method is that it requires a dyadic length which is a power of 2. To reduce the dimensionality of the training vectors, each signal was segmented into 4 equal parts. Similarly to Kadambe et al [2], we added Gaussian noise with $\sigma=0.1$ independently to the first segment of each phoneme to give an extra ten training vectors for each class. Thus we obtained a total of 42 training vectors all normalised to unit norm. The neural network classifier had 5 nodes in its hidden layer after empirically determining that this number gave sufficient classification. When

the classifier was tested on the training data it gave 100% accuracy with a 90% confidence threshold.

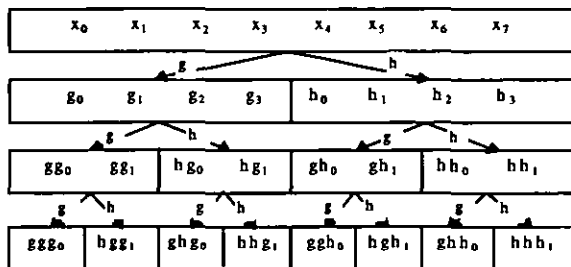


Figure 5: Best-Basis wavelet approximations organise themselves into a binary tree.

The next stage was to test the trained network on unseen data. We used the same kind of phonemes from the same speaker but uttered under different context, /d/ as in *dark*, /s/ as in *struggle*, and /z/ as in *Herb's* (see Figure 7). Overall classification was again 100% but with a lower confidence level, about 60%.

4. EVALUATION

The acoustic-phonetic feature extraction method described here takes advantage of the adaptive time-frequency localisation characteristics of the Best-Basis method to efficiently represent perceptually relevant acoustical events. That the features extracted are suitable for classification tasks has been illustrated by means of a simple training and test set consisting of those signal features contained in the wavelet coefficients. The results at this stage are promising and warrant the testing of this method on a larger database of speech data. It is interesting to note the structural similarities between the transformed data sets in Figures 6(f) and 7(a) of the contextually different phonemes /z/ used in the training and test phonemes respectively. The /s/ and /d/ phonemes exhibit a similar characteristic.

5. REFERENCES

- [1] Coifman, R.R. and Wickerhauser M.L. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [2] Kadambe, S; Srinivasan, P. "Applications of Adaptive Wavelets for Speech," *Optical Engineering* 33(7), pp.2204-2211 (July 1994).
- [3] Mallat, S.A "Theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp.674-693, 1989.
- [4] Kadambe, S. and Boudreaux-Bartels, G.F. "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- [5] Evangelista, G. "Pitch-synchronous wavelet representations of speech and music signals," *IEEE Transactions on Signal Processing*, Vol. 41, No.12, December 1993.
- [6] Evangelista, G. "Comb and multiplexed wavelet transforms and their application to speech processing," *IEEE Transactions on Signal Processing*, Vol. 42, no.2, February 1994.
- [7] Maes, S. "Nonlinear techniques for parameter extraction from quasi-continuous wavelet transform with application to speech,"

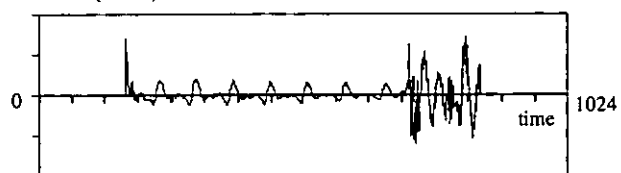
Proceedings of SPIE - The International Society for Optical Engineering 1994. Vol.2093 pp. 8-19.

[8] Szu, H; Telfer, B; Kadambe, S. "Neural network adaptive wavelets for signal representation and classification," *Optical Engineering*, vol.31 No.9 pp.1907-1916, September 1992.

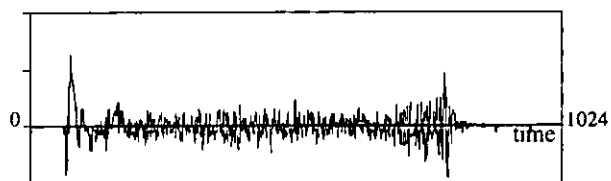
[9] Saito, N. "Local feature extraction and its application using a library of bases." *Phd thesis*, Yale University (1994).

[10] Buckheit, J.B. and Donoho, D.L. "Wavelab and reproducible research," in *Wavelets and Statistics*, Springer-Verlag, New York (1995).

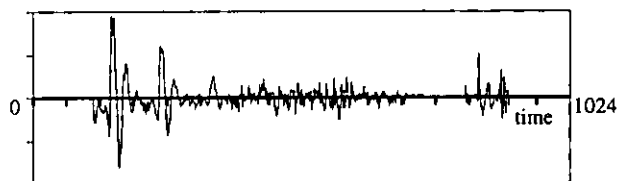
[11] Wesfried, E; Wickerhauser, M.V. "Adapted local trigonometric transforms and speech processing," *IEEE SP 41*, 3597-3600 (1993)



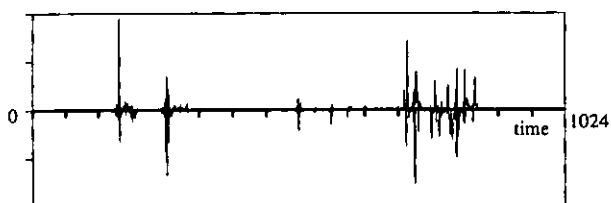
(a) /d/ as in had



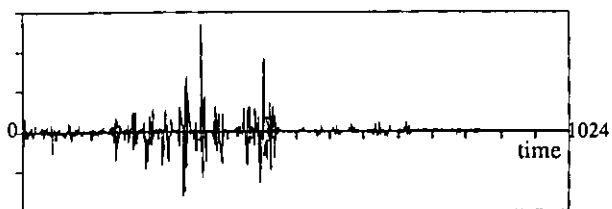
(b) /s/ as in ask



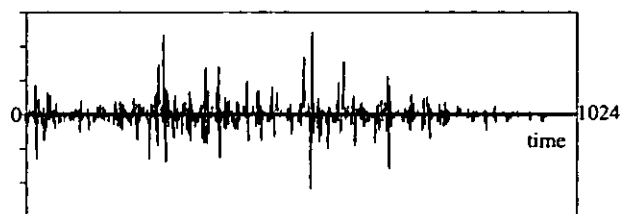
(c) /z/ as in was



(d) Transform of /d/ as in had



(e) Transform of /s/ as in ask

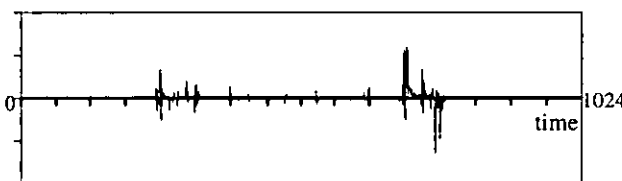


(f) Transform of /z/ as in was

Figure 6 (a)-(c): Original training signal. (d)-(f) Transform of respective signals



(a) Transform of /z/ as in Herb's



(b) Transform of /d/ as in dark



(c) Transform of /s/ as in struggle

Figure 7: Wavelet Transforms of test data. Note the correlation between transforms of contextually different phonemes.

Time-Frequency Dictionaries for Improved Discriminant Feature Extraction

C.J.Long and S.Datta

Department of Electronic and Electrical Engineering

Loughborough University of Technology

Loughborough LE11 3TU, UK.

Email c.j.long@Lboro.ac.uk

ABSTRACT

Reduction of signal dimensionality in the pre-classification stage of classification systems is usually done via one of many classical parameter extraction methods, for example Linear Predictive modelling or Fourier analysis. Many of these methods concentrate on the best possible *signal representation* and help the subsequent classification stage only in that they have effectively compressed (according to a particular criterion) the signal thus requiring less training samples.

In this paper, we consider the situation where the classification is helped in the feature extraction stage by supplying it with good *discriminative* features obtained by transformation onto a coordinate system consisting of a collection of orthonormal functions that are well localised in both time and frequency and then choosing the most suitable of these for the problem at hand.

1. INTRODUCTION

Conventional feature extraction mechanisms have typically been derived through modelling of the signal eg (LPC) or via its analysis e.g.(Fourier methods). The wavelet transform which has been recently developed is closely related to Fourier techniques but has some important fundamental differences. The two most obvious advantages are the ability of the WT to provide a multiresolutional analysis (MRA), unlike the Short Time Fourier Transform which assumes signal stationarity within its fixed time-frames, thus possibly missing important localised speech events of importance in the subsequent recognition stage. The second major difference of the WT is that the basis functions can be one of a number that can equally be used according to the problem at hand. These basis functions, or analysing functions are derived from a single *mother* wavelet and together will form an orthonormal basis of the l^2 space.

Typically, the continuous wavelet transform is overcomplete and therefore does not form

a true orthonormal basis. Redundancy may be eliminated by appropriately sampling the wavelet on a dyadic lattice, i.e. in a manner that reflects the tiling of the time-frequency plane as shown in Figure 1. An orthonormal basis of compactly supported wavelets can then be obtained to span $L^2(\mathcal{R})$, (the space of all finite energy signals) by shifting and dilating the wavelet function $\psi(t)$ i.e.

$$\psi_{n,m}(k) = 2^{-n/2} \psi(2^{-n/2} k - mb_0) \quad (3)$$

where $n=1,2, \dots$ represents the scale and $m=0,1, \dots$ the time shift. Note that the scaling factor a is here chosen as 2 in order that the frequency axis is decomposed in octaves. Now if one chooses a suitable wavelet, a true orthonormal basis will be obtained. This results in a multiresolutional analysis of a given signal over l^2 space, yielding a time -scale decomposition similar to that exhibited in Figure 1. For further details on MRA, the reader is referred to the work of Mallat [3].

In this paper, we apply the wavelet transform in the context of the Best-Basis paradigm to the problem of signal classification.

Discriminative Feature extraction is achieved via the Local Discriminant Basis

algorithm outlined in Section 2.2. The emerging basis (chosen from a library of bases) is the one that is best suited in the discriminant sense for the given problem and provides a set of Wavelet coefficients or coordinates that best highlight the differences between classes. The resulting set of features are then used to design a classifier, in our case a three-layer feed-forward neural-network upon which the efficacy of our new feature set is examined.

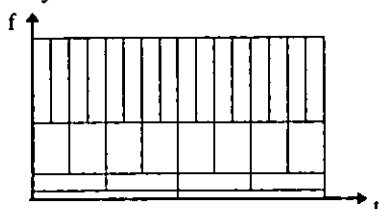


Figure 1: Tiling of time-frequency plane via the Wavelet Transform.

2. THE BEST-BASIS PARADIGM AND A LIBRARY OF BASES.

A generalisation of the Wavelet Transform originally designed for signal compression is the Best-basis algorithm first described in [1]. The idea is to do transform coding on a signal by choosing a wavelet basis which is best suited for the given problem, resulting in an adaptive time-scale analysis. In particular, two possibilities are proposed, the smooth local trigonometric transforms which essentially performs local Fourier analysis on the signal, and its frequency domain conjugate, the wavelet packet which similarly partitions the frequency axis smoothly. Since these transforms operate on recursively partitioned intervals on the respective axis, the bases whether wavelet packet or local trigonometric are said to form a *library of orthonormal bases*. If these bases are ordered by refinement, they form a tree which can be efficiently searched to result in only those coefficients which contain the most information.

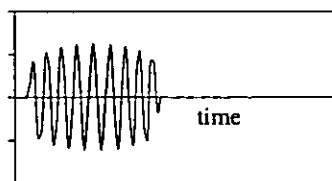


Figure 3: An example of a modulated smooth trigonometric packet. A localised sine dictionary, for example, would consist of a number of scaled, oscillatory versions of these.

In summary, the aim is to extract the maximum information or *features* from our signal by projection onto a co-ordinate system or basis function in which that signal is best (most efficiently) represented. What is meant by efficiency really depends on the final object. If compression is required, then the most efficient basis will be the one wherein most of the information is contained in just a few coefficients. On the other hand if we are interested in classification, a basis which most uniquely represents a given class of signal in the presence of other known classes will be most desirable.

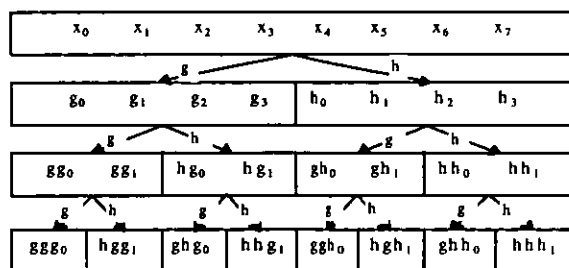


Figure 2: Best-Basis wavelet approximations organise themselves into a binary tree.

2.1 Cost Functions

The criteria used to choose the best basis via a search through the binary tree really depends on the application. Eg. For signal compression the goal is really a quest for minimum distortion and so the Shannon Entropy is a natural choice.

For the classification problem a number of measures exist see eg (4) for more detail.

Several have been suggested for use with the LDB (5) the most important of which include

- *Relative Entropy or Kullback- Leibler divergence* :

$$D(\mathbf{p}, \mathbf{q}) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right) \quad (4)$$

where the following convention holds:

$\log(0) = -\infty$, $\log(x/0) = +\infty$ for $x > 0$;

the more simple measure of

- *l² distance*:

$$D(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_2^2$$

i.e. the square of the l² norm. Note that this measure can be extended for all l^p space ($p \geq 1$).

- *Hellinger distance*:

$$D(\mathbf{p}, \mathbf{q}) = \sum_i \left(\sqrt{p(i)} - \sqrt{q(i)} \right)^2$$

The discriminant measure is used to discern which the nodes in our subspace carry the highest discriminative information and is therefore instrumental in the Best-Basis selection. The next question is what quantity should we provide to our $D(\cdot)$ (based on the Wavelet Transform) that carries all information necessary to characterise and thus discriminate amongst a collection of classes.

2.2 The Local Discriminant Basis Algorithm

In the original Best-Basis Algorithm first proposed in (1), the Shannon entropy is the measure used to search through a Wavelet Packet table similar to that shown in Figure 2. Since our problem is that of classification, we require a measure per class, based on the time-frequency dictionary that will act as a good indicator regarding the time-frequency properties of a given class. In the original LDB described in (6), a time frequency energy map is defined as a possible quantity :

$$TFE_{(j,k,m)} = \frac{\sum_{i=1}^{N_c} (w_{j,k,m} x_i^{(c)})^2}{\sum_{i=1}^{N_c} \|x_i^{(c)}\|_2^2}$$

where N_c is the number of training samples per class, j, k, m form a table with $j = 0, \dots, J$ representing depth, $k = 0, \dots, 2^j - 1$, the box index at a particular depth $m = 0, \dots, 2^{n-j} - 1$, the coefficient index within a given packet. The resulting distributions per class are then discerned between by comparing them in a pairwise manner using any of the measures in 2.1. The 'Best' Basis is selected by searching the discriminant space and choosing those subspaces (k of them) that contain the maximum amount of information (6). A *Feature Compression* stage can be subsequently carried out which chooses the top q ($< k$) most discriminant vectors.

3. EXAMPLES.

3.1 Triangular Waveform Classification

To examine the effectiveness of the LDB for extracting time-frequency features, we apply it to a three class triangular waveform problem, similar to that examined in (7). The three classes are generated according to :

$$x_1(i) = u h_1(i) + (1 - u) h_2(i) + \varepsilon(i) \text{ Class 1}$$

$$x_2(i) = u h_1(i) + (1 - u) h_3(i) + \varepsilon(i) \text{ Class 2}$$

$$x_3(i) = u h_2(i) + (1 - u) h_3(i) + \varepsilon(i) \text{ Class 3}$$

where $i = 1, 2, \dots, 32$, u is a uniform rv on (0,1) and $\varepsilon(i)$ is a noise term of normal distribution.

Training samples of 500 observations per class were generated and test samples of 300 per class. Figure 3 shows five random examples from each class.

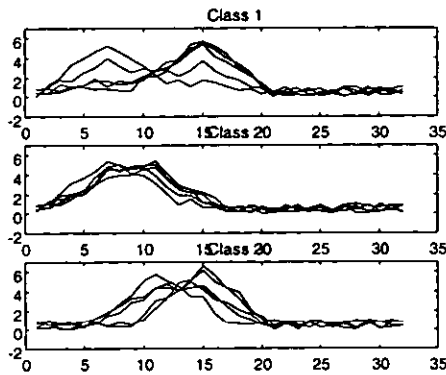


Figure 3: Waveform Data

Time-Frequency wavelet packet tables were generated using the 6-tap Coiflet wavelet packet and cosine packet tables using the smooth local trigonometric function, examples of both are shown in Figure 4.

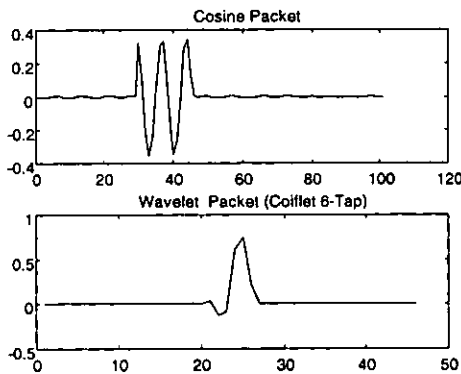


Figure 4: Some Basis Functions

In these experiments, we used all LDB features for classification and compared the performance of the technique using the three distance measures outlined in section (2.1). For a classifier we used a three layer ANN with three nodes in its hidden layer although we remark that other conventional classifiers such as CART (Classification and Regression Trees) or LDA (Linear Discriminant Analysis) may also be used. Table 1 summarizes these where we notice that LDB features derived via the Cosine Packet transform perform the overall best. The l^2 and relative entropy distance measure perform comparably well on both dictionaries with the Hellinger measure doing slightly worse. We finally note that

the LDB vectors looked similar to the h_i and their derivatives.

Technique	Error Rate (Training) %	Error Rate (Testing) %	Distance Measure
ANN on STD	20.36	22.20	-
ANN on WP (LDB32)	14.7	15.12	l^2
ANN on WP (LDB32)	15.33	16.23	Hellinger
ANN on WP (LDB32)	14.47	15.66	Relative Entropy
ANN on CP (LDB32)	12.27	14.12	l^2
ANN on CP (LDB32)	15.13	16.14	Hellinger
ANN on CP (LDB32)	12.73	13.89	Relative Entropy

4. Conclusions

In this paper we tested the LDB theory developed in (6) to a three class triangular waveform problem. Three main factors emerged:

- We looked at the effect of using different distance criteria upon the final misclassification rate, Hellinger was seen to perform marginally worse overall although the choice of distance measure is likely to be highly dependent upon the given problem.
- Two dictionaries of basis functions were used and we saw that the Cosine Packet wavelet resulted in an overall decrease in misclassification rate. This is probably due to the CP 'picking out' the temporal differences apparent amongst the classes.
- The type of subsequent classification is also critical. A simple two layer feed-forward neural network gave significantly worse performance on the LDB features in all cases except when tested on the standard Euclidean coordinates. Saito et al in (5) reports a similar depreciation in classification

when using CART as a classifier for geophysical signals and attributes this to the obliqueness of the LDB features; i.e. they must be combined linearly to reduce misclassification rates.

5. References

- 1) Coifman, R.R. and Wickerhauser M.L. "Entropy based algorithms for best-basis selection," *IEEE Transactions on Information Theory*, vol.32, pp.712-718, March 1992.
- 2) Buckheit, J.B. and Donoho, D.L. "Wavelab and reproducible research," in *Wavelets and Statistics*, Springer-Verlag, New York (1995).
- 3) Mallat, S.A "Theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp.674-693, 1989.
- 4) Basseville, M. "Distance Measures for Signal Processing and Pattern Recognition," *Signal Processing* vol.18, pp.349-369, 1989.
- 5) Saito, N. "Classification of Geophysical Acoustic Waveforms using Time-Frequency Atoms." ASA Statistical Computing Proceedings, Amer. Statist. Assoc., 1996.
- 6) Saito, N. "Local feature extraction and its application using a library of bases." *Phd thesis*, Yale University (1994).
- 7) Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984), *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Discriminant Wavelet Basis Construction for Speech Recognition

C.J.Long and S.Datta

Department of Electrical and Electronic Engineering,
Loughborough University,
Loughborough,
Leics LE11 3TU
UK

Email:C.Long@iop.bpmf.ac.uk or S.Datta@lboro.ac.uk

ABSTRACT

In this paper, a new feature extraction methodology based on Wavelet Transforms is examined, which unlike some conventional parameterisation techniques, is flexible enough to cope with the broadly differing characteristics of typical speech signals. A training phase is involved during which the final classifier is invoked to associate a cost function (a proxy for misclassification) with a given resolution. The subspaces are then searched and pruned to provide a Wavelet Basis best suited to the classification problem. Comparative results are given illustrating some improvement over the Short-Time Fourier Transform using two differing subclasses of speech.

1.1 INTRODUCTION

Multi-scale feature extraction is an attractive option when representing non-stationary real world signals such as speech. Coupled with integrated optimisation of the feature extraction and classification stages the aim is to provide an overall improvement in recognition performance. The problem is relevant because as modelling techniques have become vastly improved in recent years, further gains in recognition accuracy are likely to come from the preprocessing stage.

Wavelets and related techniques like subband coding have been applied with considerable success to speech processing applications such as compression [3],[4], and to a more limited extent on feature extraction for speech recognition / classification [5],[6].

Their main advantages are a somewhat richer multiresolution representation of the acoustic signal and the flexibility to use one of a number of basis functions. Subsequent refinements that aim to efficiently model signal statistics by choosing the depth of projection and amount of signal reduction adaptively [1] serve to improve accuracy of the model further.

Learning from the training set the best set of subspaces in which to model the data, results in a discriminant basis set which will highlight using the expansion coefficients of the wavelet transform (preferably just a few) the major differences between classes. If feature reduction is subsequently carried out, then the final classifier is designed in lower dimensional space. Assuming the data is well modelled in the first place, then there is a better chance of the classes being well separated by the classifier.

In this paper, we propose an implementation of this theoretical framework for tackling phoneme classification problems. The method is outlined in the next section.

2.1 Method

Let us first define the Discrete or Dyadic Wavelet Transform. The wavelet transform can be developed from a number of existing theories, here we will consider the extension of the DWT from its continuous counterpart; the CWT since this is intuitively similar to the Short Time Fourier Transform. The basic *analysing* or *mother* wavelet is given by:

$a^{-1/2}h(t - \tau/a)$ where τ and a are time shift and scale respectively. This shifted scaled set of functions forms an orthonormal family if sampled appropriately see [7] for further details of this. The $h(t)$ furthermore, satisfy a number of constraints to enable them to be wavelets. For example, most well designed wavelets have *compact support* both in time and frequency enabling good feature localisation in the respective domains. Wavelet *regularity*, *vanishing* moments and *orthogonality* are design parameters which influence factors such as reconstruction fidelity, degree of compression achievable, or type of signal most suitable for decomposition in that wavelet basis. A wealth of literature exists on this subject see [7], [8], [9] for details.

If we take $a = a_0^m$ and $\tau = nb_0a_0^m$, where n and m are the discretisation integers on the dyadic grid, the resulting wavelets then become

$$h_{m,n}(t) = a_0^{-m/2}h(a_0^{-m}t - nb_0) \quad (1)$$

with the added constraint that $\int h(t)dt = 0$.

The discrete wavelet transform is just a projection of a given signal onto these analysing functions:

$$c_{m,n}(f) = \langle h_{m,n}, x \rangle \quad (2)$$

The algorithm used in the following experiments is connected to the Local Discriminant Bases [1] developed for classification as a direct extension of the original Best-Basis algorithm [2]. The LDB uses dictionaries of Wavelet Packets and Local Cosine Transforms, which will be defined shortly, to form a library from which the best basis dictionary may be chosen using, as criterion, one of a number of cost functions. These cost functions, of which there are a number of differing types, are generally additive, but all essentially provide a measure of 'energy concentration' of the vector.

Definition: An additive cost function \mathcal{G}^{add} from a sequence $\{x_i\}$ to \mathcal{R} is additive if $\mathcal{G}(0)=0$ and $\mathcal{G}(\{x_i\}) = \sum_i \mathcal{G}(x_i)$.

In LDB, the cost function used is *relative entropy* which should be a good measure of the power of discrimination of each subspace. If we consider a simple two class case, where $p = \{p_i\}_{i=1}^n$, $q = \{q_i\}_{i=1}^n$ are two normalised energy distributions of signals belonging to class 1 and class 2 respectively. The Relative Entropy is then given as :

$$RE(p, q) \equiv \sum_{i=1}^n p_i \log \frac{p_i}{q_i} \quad (3)$$

Typically one first computes an estimate of the class probabilities by calculating a time-frequency energy map for each signal class from the Wavelet packet / Cosine packet transforms.

Definition: The Wavelet/Cosine Packet Transform is a generalisation of the standard discrete wavelet transform given in (2). If the signal subspace is given as $\Omega_{j,k}$ ie the coarsest resolution, then each node is split recursively in a manner similar to the DWT to form a binary tree of subnodes. If j is the depth and k the subspace number (either 0 or 1), the first level will have two subspaces, $\Omega_{j+1,k}$ and $\Omega_{j+1,k+1}$. The next will have four and the J^{th} 2^J . In total there will be 2^{2^J-1} possible subnodes in the tree and the issue is to extract a non redundant signal representation by assigning criterion such as Relative entropy to each node and pruning/growing a tree to maximise this measure.

The DWT on the other hand is iterated only on the Low Pass part of its decomposition and is such that a non-redundant representation is guaranteed. Wavelet Packets, on the other hand, have the advantage of covering signal space entirely and provide an unfixed resolution tiling of the time-frequency plane although the Heisenberg inequality principle still of course holds. However they are overcomplete and require some kind of pruning if orthogonality is to be achieved. This search will be fast if the cost function is additive.

Local Trigonometric Transforms or Sine/Cosine Packet Transforms are exactly analogous to the WP transform except that they partition the time instead of the frequency axis smoothly.

Here is the LDB algorithm used in the experiment. Assume $\Phi_{j,k}$ is the discriminant measure, whether additive or not, let $D_{j,k}$ represent the Best Discriminant Basis and $R_{j,k}$ the fully expanded, redundant basis:

- 0) Choose to use either trigonometric dictionaries or Wavelet Packets for the transform.

- 1) Expand every signal in the training set into its wavelet packet table.

- 2) Determine the set of most discriminant subspaces using a top down pruning methodology by testing the efficacy of each subspace for discrimination.

i.e. set a temporary array $\mathfrak{T}_{j,k} = \Phi_{j,k}$

if $\mathfrak{T}_{j,k} \geq \mathfrak{T}_{j+1,2k} \cup \mathfrak{T}_{j+1,2k+1}$; $D_{j,k} = R_{j,k}$;
else

$D_{j,k} = D_{j+1,2k} \oplus D_{j+1,2k+1}$ and set

$\mathfrak{T}_{j,k} = \mathfrak{T}_{j+1,2k} \cup \mathfrak{T}_{j+1,2k+1}$

- 3) Rank the expansion coefficients according to their discriminant power and from these select the top

$k \leq n$ features (where $n = 2^{no}$ is the dyadic length of the signal) for each signal in the training class to construct the final classifier.

The LDB gained from step two is an orthonormal basis, also if the cost function is additive, this step will be fast.

Step 3 isn't necessary since we can still design the classifier on all the features, however if the dimensionality of the problem is reduced, this step will reduce the number of interfering components in the decomposition, making the class-specific features more robust. Computational training times will simultaneously be reduced. In practice one can rank the expansion coefficients by a) Finding the discriminant validity of a particular basis function in the LDB expansion. b) Use Fishers class separability index to rank the coefficients.

Results

In the following experiments, the above algorithm was implemented using the standard LDB configuration: an additive cost function of Relative Entropy and the best $k \leq n$ chosen using the same criterion.

This approach was compared with a configuration using non-additive costs; a proxy for LDA-derived misclassification rate was used and the expansion coefficients ranked using Fishers class separability criterion. In addition in this case, we applied a small non-linear thresholding to the subspace vectors prior to calculating the misclassification rate. The final classifier in both cases was LDA thus in case 2 the same optimality criterion was used both in the evaluation of suitable features for class separability, as for the final classification estimate. The wavelet used in all cases was the Daubechies 6th order wavelet.

The phoneme classification problems broached dealt two extreme cases: first, three well behaved (in the statistical sense), well separated vowels aa,ax,iy corresponding to the back, mid and front positions of the tongue during voicing were examined. Secondly, the three unvoiced stops, p,t,k were discriminated against one another. In both cases, the

phonemes were extracted from dialect region 1 of the Timit database from all speakers both male and female to ensure a good statistical representation of each sound. The speech datasets used were sampled at a rate of 16Khz, thus the 32ms window, which we assumed, was composed of ~512 samples.

The results gained using the methods outlined plus a benchmark version of the STFT, commonly used in speech parameterisation are given in Table 1.

Technique		Error Rate (Training)	Error Rate (Testing)	Problem
LDA	on	9.39%	10.35%	iax
STFT64				
LDA	on	8.53%	9.40%	iax
LDB60				
LDA	on	9.2%	10.1%	iax
LDBuLDA60				
LDA	on	33.51%	43.87%	ptk
STFT64				
LDA	on	31.41%	39.68%	ptk
LDB60				
LDA	on	30.68%	42.58%	ptk
LDBuLDA60				

Table 1: Misclassification rates of the feature extraction techniques when applied to two phoneme classification problems. LDA,STFT64,LDB60 indicate the type of final classifier used, 64 short-time fourier transform gained from whole 512 via decimation, the top 60 expansion coefficients extracted using standard LDB.LDBuLDA60 is the top 60 coordinates obtained using LDA-derived optimality criterion.

CONCLUSIONS

With regard to the number of features chosen, approximately 10% of the original signal dimensionality was used. The performance of the wavelet methods was noticeably better than the STFT. The initial computational cost of the Wavelet Packet related methods is always going to be greater since there is a significant cost in the pruning part of the algorithm not present in FT methods - especially if LDA is used at this stage. However this is only a training cost, once a basis tree is worked out, all subsequent signal known to belong to a broad phonetic subclass can be decomposed in a comparably fast manner. It should also be emphasised, in particular for the ptk experiment that this is a difficult classification problem, we ourselves would generally use context and higher level knowledge to characterise these. The type of system proposed has been shown to provide some improvement over a standard widely used parameterisation technique in two situations, it is likely to be of robustly similar performance in other recognition scenarios. As a preprocessing technique to standard modelling conventions e.g. HMM it certainly shows some promise. It is likely anyway that a better recogniser would highlight improvements between Wavelet over Fourier decompositions, it has been noted in [10] that LDB derived features appeared "oblique" in a sense and this is borne out in some of our other experiments where the true

multiresolutional advantages of wavelet appeared much superior. Better performance could also be had by using some standard preprocessing of which none was done here since for the purposes of comparison this was irrelevant.

With regard to the decrease in performance between standard LDB and LDB using an LDA-derived non-additive cost, we felt was perhaps due to non-linear relations within the training set not being exploited. Instead of using LDA, in future we will try a neural network to provide a cost and incorporate this seamlessly into the whole design.

REFERENCES

- [1] N.Saito, "Local Feature Extraction and its Applications using a Library of Bases," *A Dissertation*, Yale University, Dec. 1994.
- [2] R.R.Coifman and M.V.Wickerhauser, "Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
- [3] M.V.Wickerhauser, "Acoustic Signal Compression with Wavelet Packets," In Chui C.K.(ed) *Wavelets: A Tutorial in Theory and applications* (1992).
- [4] J.A.Thiripuraneni et al, "Mixed Malvar Wavelets for Non-stationary Signal Representation," *Proc.ICASSP vol.1pp.13-16, Atlanta* (1996).
- [5] C.D'Alessandro and G.Richard, "Random wavelet representation of unvoiced speech," *Proc. IEEE-SP Int.Symp. on Time Frequency and Time Scale Analysis*, pp. 41-44 (Oct.1992).
- [6] S.Kadambe and G.Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," *IEEE Transactions on Information Theory*, vol. 38, no.2, pp. 713-718, March 1992.
- [7] I.Daubechies, "Orthonormal bases of compactly supported wavelets," *Comm in Pure and Applied Math.*, vol.41 No.7, pp.909-996, 1988.
- [8] S.Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, pp. 674-693, 1989.
- [9] O.Rioui and M.Vetterli, "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*, October 1991.
- [10] N.Saito, "Classification of Geophysical Acoustic Waveforms using Time-Frequency Atoms," *Proceedings of Statistical Computing, Amer. Statist. Assoc.* (1996).

