The investigation into an algorithm based on wavelet basis functions for the spatial and frequency decomposition of arbitrary signals

by

Hilton Goldstein

Submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

in the

Department of Computer Science

University Of Natal, Durban

January 1994

Abstract

The research was directed toward the viability of an $O(n)$ algorithm which could decompose an arbitrary signal (sound, vibration etc.) into its time-frequency space. The well known Fourier Transform uses sine and cosine functions (having infinite support on t) as orthonormal basis functions to decompose a signal $f(t)$ in the time domain to $F(\omega)$ in the frequency domain, where the Fourier coefficients $F(\omega)$ are the contributions of each frequency in the original signal. Due to the non-local support of these basis functions, a signal containing a sharp localised transient does not have localised coefficients, but rather coefficients that decay slowly. Another problem is that the coefficients $F(\omega)$ do not convey any time information. The windowed Fourier Transform, or short-time Fourier Transform, does attempt to resolve the latter, but has had limited success.

Wavelets are basis functions, usually mutually orthonormal, having finite support in t and are therefore spatially local. Using non-orthogonal wavelets, the Dominant Scale Transform (DST) designed by the author, decomposes a signal into its approximate time-frequency space. The associated Dominant Scale Algorithm (DSA) has $O(n)$ complexity and is integer-based. These two characteristics make the DSA extremely efficient. The thesis also investigates the problem of converting a music signal into it's equivalent music score. The old problem of speech recognition is also examined. The results obtained from the DST are shown to be consistent with those of other authors who have utilised other methods. The resulting DST coefficients are shown to render the DST particularly useful in speech segmentation (silence regions, voiced speech regions, and frication). Moreover, the Spectrogram Dominant Scale Transform (SDST), formulated from the DST, was shown to approximate the Fourier coefficients over fixed time intervals within vowel regions of human speech.

Acknowledgements

I would like to thank my friends and colleagues who have made significant contributions, both on the academic and personal side (in no particular order):

- Professor A.G. Sartori-Angus for being my supervisor.

- Hugh C. Murrell for introducing me to some of the theory and applications of wavelets and for being available to answer questions.

- Vivienne Smithdorf for the superb effort she put into proof-reading my final manuscript. Towards the end of my write-up, when I was under the most pressure, it really was gratifying to have someone so willing and capable as Vivienne to take a huge load off my shoulders. I also find it very comforting that I have someone to blame for any spelling or grammatical errors! Thanks Viv!

- David Ian Carson and Angela Kay Cooper for volunteering their time to digitally record practically hundreds of speech samples.

- Professor M.J. Alport in the Department of Physics at the University of Natal (Durban) for encouraging me to examine some tough physics problems. I feel that inter-disciplinary research is probably the most satisfying of all.

- Dave Fraser in the Department of Electronic Engineering at the University Of Natal (Durban) for the valuable information pertaining to signal phase content.

- My mother and father without whose support, in all forms from age zero to the present, I certainly could not have achieved what I have. Appreciation of any nature would simply pale into insignificance next to the contribution you have made to my life and education, but I offer it, nevertheless.

- My brother Colin in San José, CA (United States of America) for all the help with file formats, the endless chats on InterNet, and steady encouragement throughout the research and write-up. Thanks Col!

Acknowledgements (continued)

- Steven Adrian West for generating the data sets for several wavelet basis functions and their associated scaling functions. Also, for the tireless effort put into producing some of the illustrations appearing in the thesis. I really appreciated your encouragement throughout my write-up.

- To the boys and girls from Altyr Ego. Mike and Tania, Dave and Jackie, Gav and Trace, and Wayne and Vicki - the hours we spent together and the 'start/stop' pressure I was put under every Thursday night, made the last year of my research particularly enjoyable. Thanks guys...

- Last, but certainly not least, to my friends. Just your being my friend is very important to me - thank you...

Preface

All research described in this thesis was carried out in the Department of Computer Science, University of Natal, Durban from April 1991 to December 1993 under the supervision of Professor A. G. Sartori-Angus.

These studies represent original work by the author and have not been submitted in part or in whole to this or any other University.

Table Of Contents

Chapter 1

Introduction

1.1 THE CONCEPT OF FREQUENCY ANALYSIS

The *frequency domain* is the domain in which a signal's amplitude is examined as a function of frequency. Fourier analysis or frequency analysis is, in the simplest sense, the study of the effects of adding together sine and cosine functions. This type of analysis has become an essential tool in the study of a remarkably large number of engineering and scientific problems. Daniel Bernoulli, while studying vibrations of a string in the 1750s, first suggested that a continuous function over the interval (0, π) could be represented by an infinite series consisting only of sine functions. Bernoulli's suggestion was based on his physical intuition and was not readily accepted by mathematicians. Roughly 70 years later J.B. Fourier re-opened the controversy while studying heat transfer. He argued, more formally, that a function continuous on an interval (-π, π) could be represented as a linear combination of both sine and cosine functions. However his conjecture was not readily accepted, and the question went unresolved for many years [Wea83].

Nowadays, Fourier Transforms play an important part in many scientific disciplines. With the advent of the Fast Fourier Transform (FFT) in the mid-1960s [Coo65], as well as the birth of powerful low-cost personal computers, the spectral analysis of large data sets became practical. The FFT eliminates redundant calculations which occur in the Fourier Transform, thereby reducing the computational complexity from $O(n^2)$ to $O(n \log_2 n)$. Image and signal processing, Computed Tomography (CT) scans, vibrational systems, electronic circuitry, speech recognition, and optics are just a few fields where the Fourier Transform has become an essential tool. An oscilloscope enables us to see the shape of an electrical waveform, and a spectroscope or spectrum analyser enables us to see optical or electrical spectra. Our acoustical appreciation is even more direct, since the human ear hears spectra; see Chapter 3. Waveforms and spectra are Fourier transforms of each other [Bra65].

1.2 SOME DEFINITIONS

We start with some definitions that are used throughout this text.

Definition: A sequence $f(t)$ is said to have *finite support* if it is zero outside some region of the domain of $f(t)$, called its *region of support.* The *length* of a sequence is defined to be the size of its region of support.

Definition: A sequence $f(t)$ is *periodic,* with period *N*, if $f(t+kN)=f(t)\ \forall\ k \in Z$.

Definition: $\delta_{mn}=1$ iff $m=n$, 0 otherwise; δ_{mn} is known as the Kronecker delta.

Definition: Two functions are orthogonal iff $\int f^* g\, dt = k\delta_{mn}$ where f^* is the complex conjugate of f and $k \in \Re$. Orthonormality holds iff $k=1$.

Definition: For $l^2(Z)$, the set of all square summable sequences of complex numbers indexed by integers, we have:

$$\langle f, g\rangle = \sum_i f_i\, \overline{g}_i$$

and for non-complex f and g, simply:

$$\langle f, g\rangle = \sum_i f_i\, g_i$$

Definition: The sampling frequency of f will be denoted by f_s.

1.3 THE FOURIER TRANSFORM

The Fourier Transform, introduced by J.B. Fourier (1768-1830), is the mathematical basis of frequency analysis. Given a function $f(t)$, we define the *Fourier Transform* and the *Inverse Fourier Transform* pair as:

$$F(\omega) \quad = \quad \frac{1}{\sqrt{2}}\int f(t)e^{-i\omega t}dt \qquad (1\text{-}1)$$

and

$$f(t) = \frac{1}{\sqrt{2}} \int F(\omega) e^{i\omega t} d\omega \qquad (1\text{-}2)$$

The Fourier Transform uses sine and cosine functions as orthonormal basis functions to decompose a signal $f(t)$ in the time domain to $F(\omega)$ in the frequency domain, where the Fourier coefficients $F(\omega)$ are the contributions of each frequency in the original signal. The sine and cosine basis functions are orthogonal and local in frequency, but global in t. This accounts for two major disadvantages of the Fourier Transform. Due to the non-local support of these basis functions, a signal containing a sharp localised transient does not have localised coefficients, but rather coefficients that decay slowly. Another problem is that the coefficients $F(\omega)$ do not convey any time information. The windowed Fourier Transform, or *Short-Time Fourier Transform* (STFT), does attempt to resolve the latter problem, but has had limited success [Ran87]. The Fourier Transform is an $O(n^2)$ transform. Equations 1-1 and 1-2 can be computed considerably faster using the Fast Fourier Transform (FFT) algorithm originally designed by Cooley and Tukey [Coo65]. The Fast Fourier Transform is an $O(n \log_2 n)$ transform. A huge array of literature, both theoretical and applied, has been published on the Fourier Transform. The *Fast Hartley Transform* (FHT), which is also an $O(n \log_2 n)$ transform, proves to be slightly more efficient than the FFT, while still preserving the FFT's properties [Bra83, Mur90].

In practice, computers interpret functions as a sequence of numbers rather than a continuous function. Therefore, given any bounded N^{th} order sequence $f(t)$, the *Discrete Fourier Transform* and the *Discrete Inverse Fourier Transform* pair is defined as:

$$F(\omega) = \frac{1}{N} \sum_{t=0}^{N-1} f(t)\, e^{-2\pi\omega t/N} \qquad (1\text{-}3)$$

and

$$f(t) = \sum_{j=0}^{N-1} F(j)\, e^{2\pi tj/N} \qquad (1\text{-}4)$$

The Fourier Transform is a linear transform which takes, as its input, a signal $f(t)$ of length n in the time domain (assumed to be periodic) and produces, as its output, a signal $F(\omega)$ of length n in the frequency domain consisting of Fourier coefficients. Each of these Fourier coefficients represents a specific frequency in the original signal. For signals in the discrete time domain, the *Discrete Fourier Transform* (DFT), or the more computationally efficient Fast Fourier Transform, is used.

In many applications, the components of the resulting vector are complex numbers in which we are interested only in the magnitude, given as usual by $\sqrt{\mathrm{Re}^2(x)+\mathrm{Im}^2(x)}$. The real and imaginary components of each spectral coefficient can be used to determine the phase of its corresponding frequency component. However, in most cases, the phase localisation is unimportant. More importantly, the factor $e^{i\omega}$ is equivalent to $\sin(\omega)+i\cos(\omega)$ which has the consequence that the FFT attempts to construct the original signal from sinusoidal components, just as Bernoulli had previously speculated. Its decomposition of non-sinusoidal signals often leads to the summing over many terms of rapidly varying phase. The forward transform discards direct reference to time and the resulting coefficients are representative of the entire signal. This has two consequences; namely, the transform is not time-localised, and short-time transients may significantly affect many coefficients not corresponding to that of the transient. Another characteristic of the FFT is that the original signal is assumed to be periodic; furthermore, the frequency scale is linear. Figures 1-1 and 1-2 show a synthesised periodic signal from an organ and its corresponding Fourier coefficients.

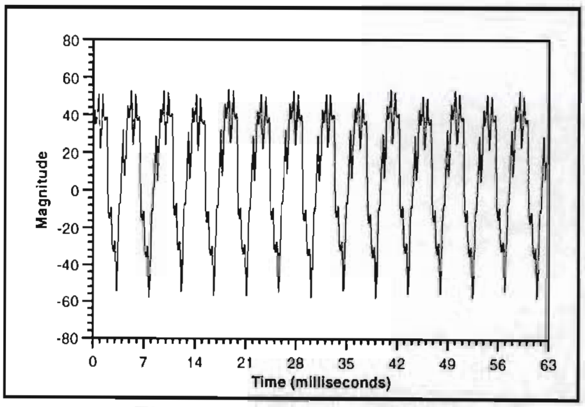


Figure 1-1: Organ signal.

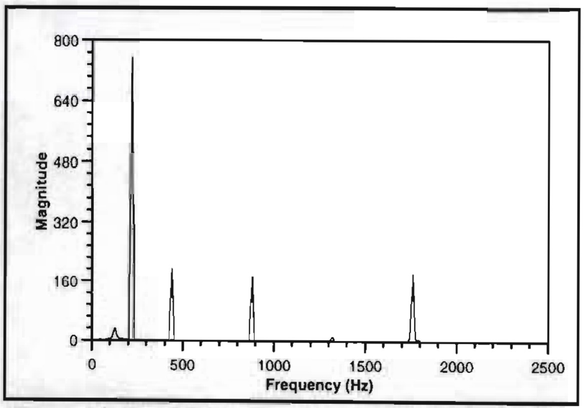


Figure 1-2: Fourier Transform of organ signal.

As another example, consider vibrations of a motor engine being measured at a predetermined speed in RPM where the fundamental frequency as well as the harmonics would remain relatively constant for long periods of time. This scenario suits the Fourier Transform which assumes the signal has a periodic nature.

1.4 Shortcomings of the Fourier Transform

The assumption of a periodic signal cannot always be made, and when periodicity is not present, the Fourier Transform's performance is often less than expected. We now examine the following examples:

- An *impulse,* by which we mean an extremely brief, very intense pulse.
- A sine wave which has increasing frequency, and is therefore *time-dependent.*
- A zero-padded signal allowing the $O(n \log_2 n)$ Fast Fourier Transform to be used instead of the slower $O(n^2)$ Fourier Transform.
- A pure sine wave which undergoes a π radian *phase change.*

1.4.1 The Impulse

Noise is the contribution to a signal which causes a deviation from the expected signal. The infinite-support basis functions cannot elegantly be used to decompose a signal affected by noise. Consider the following example. A pure sine wave was generated and its Fourier coefficients generated. Thereafter, a single spike, characteristic of "salt-and-pepper noise", was superimposed onto the sine wave and the Fourier coefficients re-calculated. The degree to which a spike or "salt-and-pepper noise" is classed as high-energy is relative to the signal upon which it is superimposed. As we use almost the full amplitude range later in the thesis to record the speech samples, such a high-energy spike would not be characteristic. Since the recording was performed under controlled conditions, the amplitude range could be chosen with much flexibility, thereby maximising the recording range. We used a spike having very high energy (100× signal amplitude) to illustrate the dramatic effect a single element can have on every coefficient in $F(\omega)$; see Figures 1-3 and 1-4.

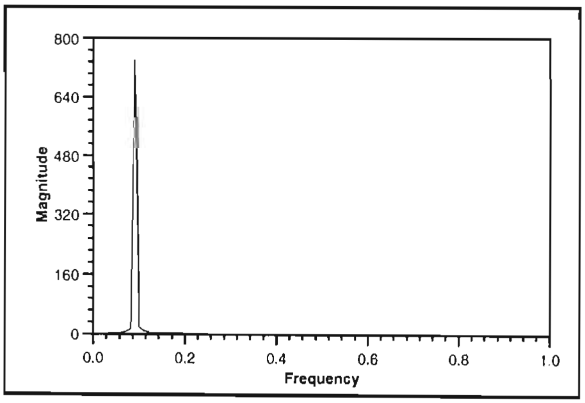


Figure 1-3: FFT of sine wave.

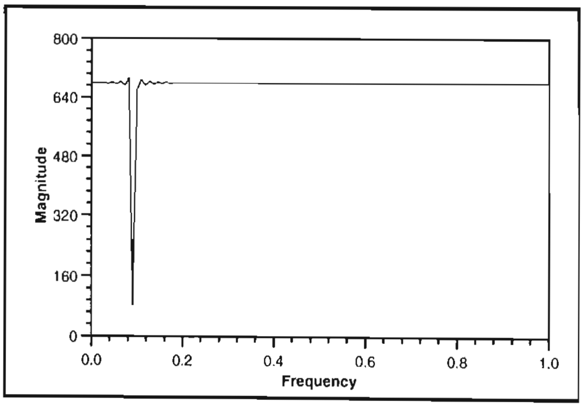


Figure 1-4: FFT of sine wave plus spike.

1.4.2 Frequency time-dependence

The infinite-support basis functions of the Fourier Transform do not lend themselves to the localisation of the signal $f(t)$. Many signals do not exhibit a periodic nature, but rather have many frequency transients. Moreover, the spatial information of these frequencies often becomes important. The sine and cosine basis functions are orthogonal and local in frequency, but global in *t*. Therefore, the Fourier coefficients $F(\omega)$ do not, in their native state, convey any time information. The forward Fourier Transform in equation 1-3 gives us a representation of the frequency content of $f(t)$, but information concerning time-localisation cannot easily be determined from $F(\omega)$.

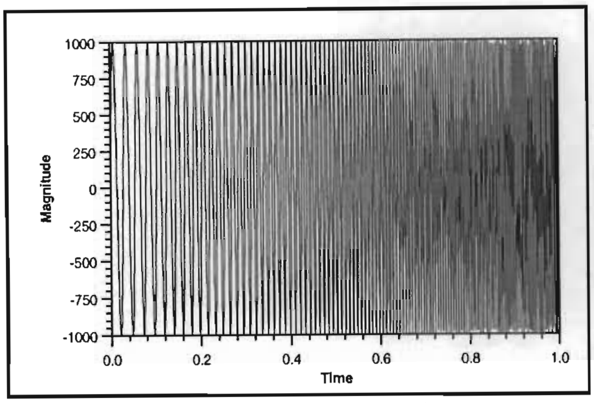


Figure 1-5: Sine wave with linearly increasing frequency.

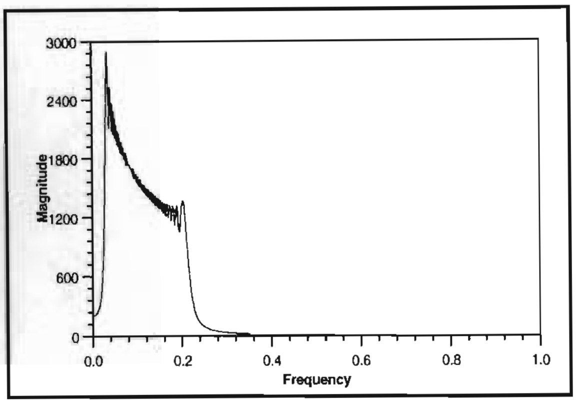


Figure 1-6: Its corresponding Fourier Transform.

1.4.3 Length of power of 2

There exist Fourier Transforms for data sets of length *N* not a power of 2. They subdivide the initial data set into successively smaller data sets, not by a factor of 2, but by whatever small prime factors divide *N*. For some specially favourable values of *N*, the Winograd algorithms can be significantly faster than the simpler FFT algorithms. This advantage in speed, however, must be weighed against the considerably more complicated data indexing involved in these transforms [Pre88]. For the simpler FFT algorithm to be used on a data set not of length of power of 2, the signal must be padded, usually with zeros, to obtain a length of 2^n. As an example, a sinusoidal signal of length 600 was generated and its Fourier coefficients calculated. The signal was then padded with zeros and its Fourier coefficients re-calculated. The Fourier coefficients for both are shown below. Notice that only slight changes in the coefficient set were observed. This could be viewed as a windowed Fourier Transform where the window in this case is rectangular. The difference between the two plots can be examined in more detail by examining the Fourier coefficients of the rectangular window (Convolution Theorem).

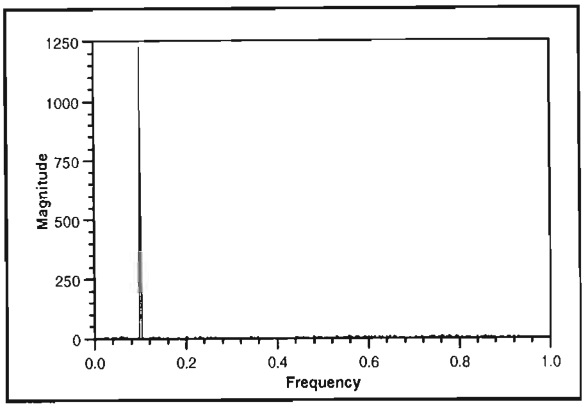


Figure 1-7: Fourier coefficients.

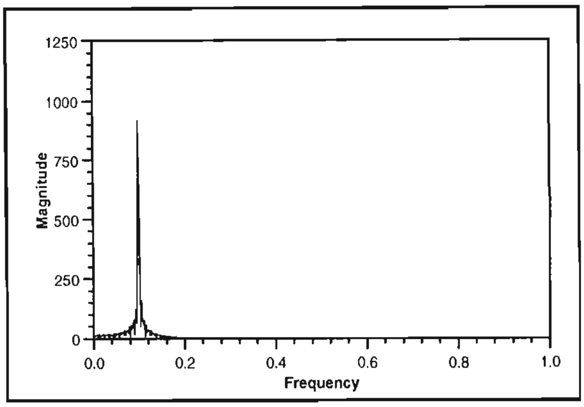


Figure 1-8: Fast Fourier coefficients.

1.4.4 Phase changes

Signals can undergo phase changes due to a variety of phenomena. How these phase changes affect their spectra can be important. Phase information can also be extremely valuable in many applications. For example, phase-shift keying is of prime importance in communications systems. We now examine the interesting effect that a phase change has on the Fourier coefficients. The introduction of a $\pi/2$ radian phase change in a pure cosine wave results in the spectrum shown in Figure 1-10. Note that we are concentrating on only one-eighth of the frequency spectrum. The spectrum is slightly distorted from its anticipated single sharp peak.

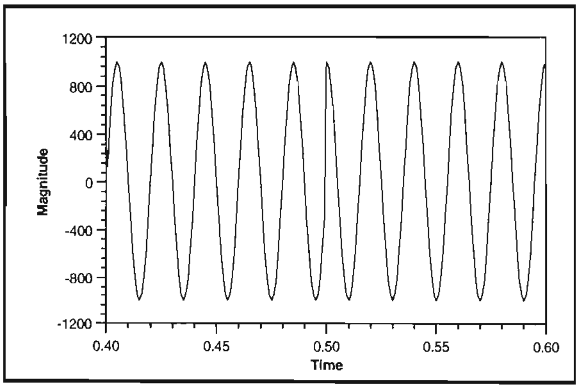


Figure 1-9: Sine wave undergoing a $\pi/2$ radian phase change.

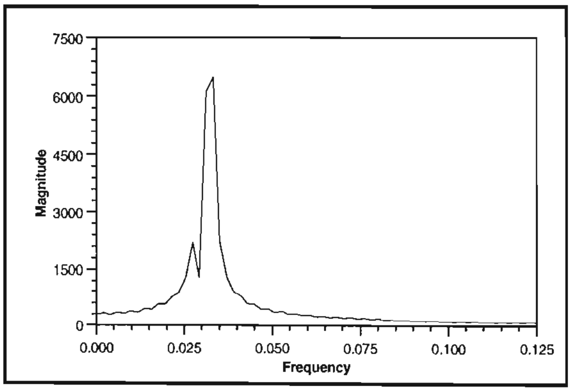


Figure 1-10: Fourier coefficients.

The introduction of a π radian phase change at an extremum can dramatically affect the Fourier coefficients. The time-domain and frequency-domain plots are shown in Figures 1-11 and 1-12. Note that there are now two spectral peaks present.

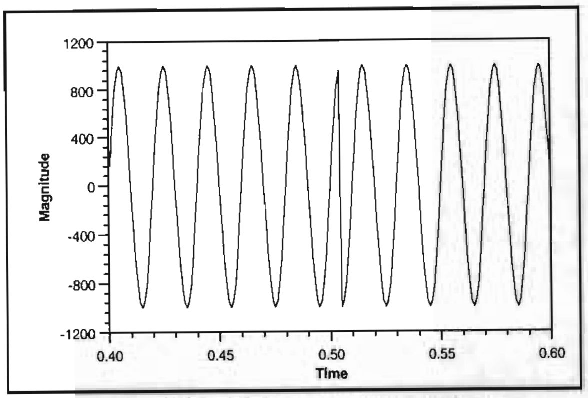


Figure 1-11: Sine wave undergoing a π radian phase change.

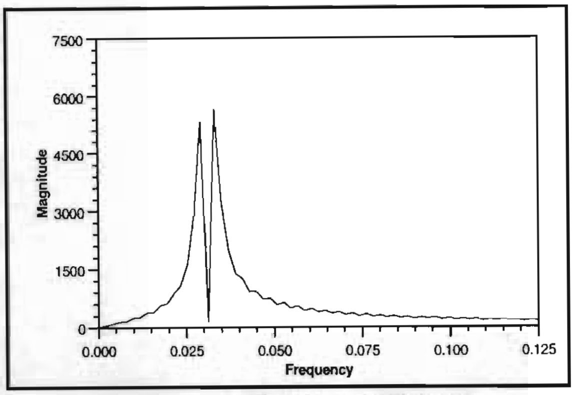


Figure 1-12: Fourier coefficients.

A Hamming window [Cun92] was applied to the time-domain signal before the Fourier Transform was effected. The Hamming window is given by:

$$w_H(n) \quad = \quad \begin{cases} \alpha + (1-\alpha)\cos\frac{\pi n}{N} & |n| \le N \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } \alpha = 0.54 \qquad (1\text{-}5)$$

1.5 Direction of the Following Text

This first chapter has briefly introduced the notion of frequency analysis and also acquainted the reader with some of the shortcomings of the well-established Fourier Transform. The research carried out and documented herein was not an attempt to further the mathematical base of knowledge on wavelets. Rather, the focus of the research was to attempt to furnish an efficient algorithm with characteristics similar to the Fourier Transform, but one which does not suffer its shortcomings, some of which we have just seen. Due to the practical nature of the research, most of the results appearing in the following pages are in the form of graphs.

1.5.1 Thesis Structure and Contribution of the Research

The research carried out by the author has resulted in an integer-based $O(n)$ algorithm for the decomposing of arbitrary signals into their time-frequency space. The algorithm has been called the *Dominant Scale Algorithm* (DSA) and the resulting time-frequency coefficients are called the *Dominant Scale Transform* (DST) for reasons which shall become apparent later. The examples presented in this chapter will be used again later to illustrate how the DST overcomes the problems described earlier. Chapter 2 introduces families of finite-support basis functions which are dilations and translations of each other.

This type of basis function is known as a *wavelet*. Orthogonal wavelets allow for the rapid decomposition of signals and the *Fast Wavelet Transform* (FWT) from Mallat [Mal89b] is described with a view to spectral analysis. The DST and associated DSA are also described. The algorithm forms the foundation of the research presented in the thesis. We document various properties of the DST. Chapter 3 tackles the complex problem of speech recognition, and as this field poses a long standing problem in both the computer science and mathematics fields, it is well studied. This allows us to verify the consistency of our results with other authors in the field. We show that by simple manipulation of the DST, the *Spectrogram Dominant Scale Transform* (SDST) yields coefficients which are remarkably similar to those of the Fourier Transform. For this, we chose 270 human speech signals generated by three individuals (two males, one female) upon which to base our research. The chapter examines how the SDST can be used as an approximation to the Fourier coefficients for vowel selection in human speech recognition. Chapter 4 examines various other time-dependent real-time problems and the DST's applicability to solving them. The problems examined are: music decomposition, rotational velocity of a projectile under extreme acceleration, and periodic pulling [Las69, Koe93]. Chapter 5 concludes this thesis by reviewing the contribution of the research as well as providing the reader with a view to envisaged future research.

Please note: Before considering Chapter 2, the reader is encouraged to read Appendix A which details the hardware, software, and sampling processes used during the research throughout the thesis.

Chapter 2

Time-Frequency Decomposition and The Dominant Scale Transform

2.1 INTRODUCTION

For many different types of signals, much of the important information carried by the signals is conveyed by singularities and transients. Examples of situations concerned with such signals are: detection of anomalies in heart beats, the analysis of vibrations in vehicles, the study/recognition of human speech, measuring rotational velocities of projectiles, and the creation of a music score from a music signal wave. We shall investigate some of these applications in Chapters 3 and 4. In two-dimensional signals, we note that the sharp variation points provide the locations of contours in images. These contours are often the most important image features and the location of these contours is a well-known problem in image-processing known as *edge detection.* Edge detection occurring in the eye enhances our ability to recognise objects from a drawing that only outlines edges. Although our entire discussion in this text is based on one-dimensional signals, we propose that many of the techniques described herein could be applied to two-dimensional signals, such as 2D images, by treating each row or column of the image as a separate entity.

This chapter briefly examines the established Short-Time Fourier Transform and the relatively new Wavelet Transform which is used when dealing with *non-stationary signals.* Non-stationary signals are those signals in which there exist frequency transients, or short-lived oscillations. Thereafter, the new Dominant Scale Transform (DST), which approximates a signal in the time-frequency space, is introduced. Properties of the DST are discussed. First, however, we briefly examine established local spectral decomposition methods.

2.2 The Short-Time Fourier Transform

In signal analysis, one often encounters the so-called *Short-Time Fourier Transform* (STFT), or windowed Fourier Transform. This consists of multiplying the signal $f(t)$ with a usually compactly supported window function $g(t)$ centred around zero, and then computing the coefficients of the product gf. These coefficients give an indication of the frequency content of the signal $f(t)$ in the neighbourhood of $t = 0$. This procedure is then repeated with translated versions of the window function, i.e., $g(t \pm t_0)$, $g(t \pm 2t_0)$, $g(t \pm 3t_0)$... where t_0 is a suitably chosen time translation step. This results in the collection of Fourier coefficients:

$$c_{mn}(f) \quad = \quad \int f(t)\, g(t - nt_0)\, e^{im\omega_0 t} dt \qquad (2\text{-}1)$$

Time-localisation is achieved by first applying a windowing mask $g(t)$ to the signal $f(t)$, thereby isolating a relatively well-localised portion of $f(t)$, and then taking its Fourier Transform:

$$F^{win}(\omega, t) \quad = \quad \int f(s)\, e^{-i\omega s} g(s - t) ds \qquad (2\text{-}2)$$

It is even more familiar to signal analysts in its discrete version, where t and ω are assigned regularly spaced values: $t = nt_0$, $\omega = m\omega_0$ where m, n range over Z, and $\omega_0, t_0 > 0$. Then Eq. 2-2 becomes:

$$F^{win}_{m,n}(f) \quad = \quad \int f(s)\, e^{-im\omega_0 s} g(s - nt_0) ds \qquad (2\text{-}3)$$

The *Windowed Fourier Transform*, given in Eq. 2-3, is a standard technique for time-frequency localisation [Ran87, Dau92]. A characteristic of the STFT is that, for time-frequency localisations, the time-frequency window has constant size at all frequencies. This inflexibility of the STFT restricts its range of applications in the study of non-stationary signals having wide frequency ranges. Choosing the length of the windowing function $g(t)$ involves calculated trade-offs:

- *Long STFT windows* provide good frequency resolution, but poor time resolution. Spectral bleeding increases proportionally to the length of the window. High frequency signals require narrow time-windows for spatial accuracy.

- *Short STFT windows* provide poor frequency resolution, but good time resolution. Short windows enable the optimal localisation of features, but may not have sufficient length to encompass the lower frequencies. Low-frequency signals require wide time-windows for studying complete cycles.

This results in a compromise in the length of the window used in short-time analysis. For example, in speech processing, the window length is dependent on the pitch ranges between a male and a female or child. The choice of the shape of $g(t)$ relies primarily on its Fourier Transform (Convolution Theorem). Some popular choices for $g(t)$ are the Rectangular, Hamming, Hanning, Kaiser, and Lanczos windowing functions. The use of windowing functions, as well as problems introduced by the Windowed Fourier Transform is documented in [Wea83, Fal85].

2.3 Established Time-Frequency Localisation Techniques

2.3.1 The Gabor Transform

The concept of the *Wavelet Transform* was initially devised by a French geophysicist, Jean Morlet [Gou84] to use in high-resolution seismic methods in oil and gas field development. The tests involved the use of back-scattered energy rather than that of reflected signals. Representations of seismic traces in the time-frequency domain were of interest. The deficiencies of the Fourier Transform were observed by D. Gabor, and in his 1946 paper he introduced a time-localisation window function $g(t-b)$ where the parameter b translates the window to cover the entire time-domain, thereby extracting spatially localised frequency content. Gabor's representation [Gab46] was based on a family of two-parameter basis functions in which all basis functions are shifts in time and frequency of the others. The original proposal used a Gaussian function g and parameters ω_0 and t_0 such that $\omega_0 t_0 = 2\pi$. One property of the function is that it is optimally concentrated in both time and frequency, and therefore well suited for an analysis in which both time and frequency localisation are important. Unfortunately, the original proposal of $\omega_0 t_0 = 2\pi$ leads to unstable reconstruction. A short, high frequency transient in a signal results in the summing of many sinusoidals having varied phase, and therefore depends heavily on cancellation. This leads to instability in the numerical calculations [Kro87]. The function

$g_{\omega_0,t_0}(t) = e^{-i\omega_0 t} g(t - t_0)$ can be viewed as translated envelopes g, "filled in" with higher frequencies.

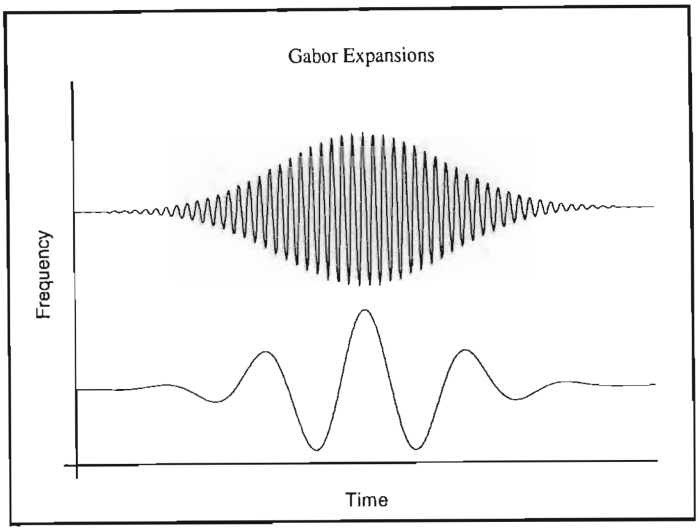


Figure 2-1: Gabor expansions in time-frequency domains.

2.3.2 The Wavelet Transform

The name *wavelet* was coined approximately a decade ago [Mor82, Mor83, Gro84] and, in the last ten years, interest has grown at an explosive rate predominantly in applied mathematics and signal processing. There are several reasons for the present success of wavelets. On the one hand, the idea of wavelets can be viewed as a synthesis of ideas which originated during the last twenty or thirty years in engineering, physics, and pure mathematics. As a consequence of these inter-disciplinary origins, wavelets appeal to scientists and engineers of many different backgrounds. On the other hand, wavelets are a fairly simple mathematical tool with a great variety of possible applications.

There has recently been a significant amount of research into wavelet transforms by various mathematicians: Yves Meyer [Mey86], Ingrid Daubechies [Dau88a, Dau88b, Dau90, Dau92], and Stephane Mallat [Mal89a, Mal89b, Mal89c], among others. Already wavelets have led to applications in signal analysis [Kro87] and numerical analysis [Bey91]; many other applications are being studied. Wavelet theory is related to Quadrature Mirror Filters that are used in image compression, progressive transmission, orientation analysis, motion analysis, and computer vision [Woo86, Ade87, Mal89a, Mal89c, Sim90]. Real-time wavelet-based video compression techniques are achieving promising compression ratios [Lew89]. Wavelet theory has even been applied to the study of the galaxy distribution [Sle90]. Active sonar applications have been developed [Fla90]. Integrated circuit manufacturers have started building chips which implement the wavelet transform entirely in silicon [Kno90].

The term "wavelets" refers to families of functions of the form:

$$\psi_{a,b}(t) \quad = \quad |a|^{-1/2}\psi\left(\frac{t-b}{a}\right) \tag{2-4}$$

which describes a two-parameter family of basis functions formed by *translations* and *dilations* of a single function $\psi(t)$ which is called the *analysing wavelet, basic wavelet*, or *mother wavelet*; we shall use the latter term throughout the text. Dilations replace the frequency translations of the Gabor expansions. Wavelets are the building blocks of wavelet analysis just as the sinusoidal functions are the building blocks of Fourier analysis. Wavelet basis functions, in contrast to the Gabor expansions, are translations and dilations of the other basis functions. Obviously, the translation does not affect the shape of the wavelet in any manner. However, it is interesting to note that the dilation process does not alter the number of cycles as in the Gabor process; rather, the extent of the window varies.

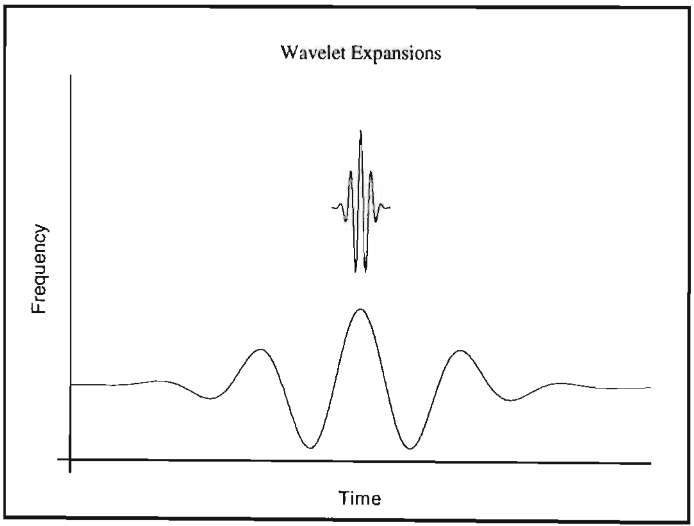


Figure 2-2: Wavelet expansions in time-dilation domain.

Note that in Figure 2-2, we have taken the liberty of representing dilations as changes in frequency. Since the frequency of a signal is proportional to the length of its cycle, it follows that for high-frequency spectral information, the time-interval should be smaller than that for low-frequency spectral information. In other words, it is important to have a window which is flexible, and narrows at high-frequencies and widens at low-frequencies. This characteristic is referred to as *zooming*. It is this zooming characteristic that allows wavelets to detect, and clearly represent, transients within a signal. The Fourier Transform has an inverse (see Eq. 1-2); therefore, the Fourier coefficients will contain inherent information about any transients. This information is, however, not easily extractable.

2.3.2.1 The Scaling Function

Devising basis functions which have local support in *t*, but which still maintain mutual orthogonality, is examined in [Dau88a, Mal89a, Str89]. Wavelets have fixed shape and are based on translations (shift in *t*) and dilations (expansion in *t*) of a mother wavelet. More specifically, wavelets constitute a family of functions, each a translation and dilation of the others, derived from one function. Construction of a mother wavelet begins with a *scaling function* ϕ. To generate this scaling function, we start with a box function ϕ_0:

$$\phi_0(t) \quad = \quad \begin{cases} 1 & \text{for } 0 \le t < 1 \\ 0 & \text{otherwise} \end{cases} \qquad (2\text{-}5)$$

To create the scaling function, we choose coefficients $\{c_k\}$ and iterate:

$$\phi_j(t) \quad = \quad \sum c_k \phi_{j-1}(2t-k) \qquad (2\text{-}6)$$

The result tends to the required scaling function, defined as:

$$\phi \quad = \quad \lim_{j \to \infty} \phi_j \qquad (2\text{-}7)$$

To normalise the scaling function, we require that $\int \phi \, dt = 1$. So using Eq. 2-6, we have:

$$\int \phi \, dt \quad = \quad \tfrac{1}{2} \sum c_k \int \phi(2t-k) \; d(2t-k) \qquad (2\text{-}8)$$

It follows therefore, that a constraint on the choice of the coefficients is $\Sigma \, c_k = 2$. Both recursive and iterative methods of constructing wavelets are documented in [Str89]. Examples of scaling functions are shown in Figure 2-3. These are the box function and hat function with coefficients $\{1 \; 1\}$ and $\{\tfrac{1}{2} \; 1 \; \tfrac{1}{2}\}$ respectively.

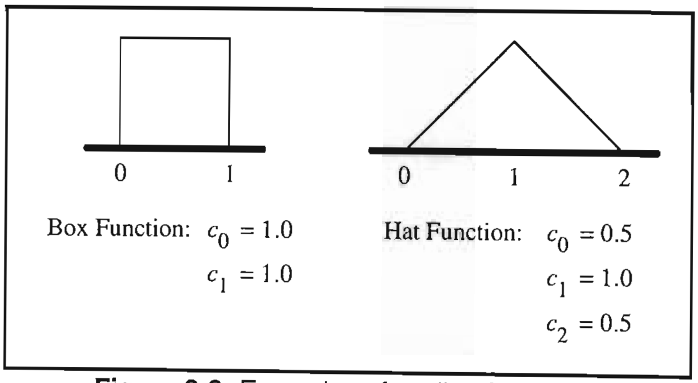


Figure 2-3: Examples of scaling functions.

2.3.2.2 Construction of the Mother Wavelet

From the scaling function, we construct the mother wavelet using the same coefficients $\{c_k\}$ according to:

$$\psi(t) \quad = \quad \sum_k (-1)^k c_{k+1} \phi(2t+k) \qquad (2\text{-}9)$$

We have great latitude in our choice of the mother wavelet. However, certain well-defined conditions apply [Gro84] and we therefore cannot choose it arbitrarily. Let $\psi(t)$ be a mother wavelet. The following conditions must hold:

1. $\psi(t)$ is absolutely integrable and square integrable. The latter condition means that it must have finite energy.

$$\int \psi(t)\, dt < \infty \text{ and } \int |\psi(t)|^2 < \infty \qquad (2\text{-}10)$$

2. If $\hat{\psi}(\omega)$ represents the Fourier Transform of $\psi(t)$, then:

$$\int \frac{|\hat{\psi}(\omega)|^2}{\omega}\, d\omega < \infty \qquad (2\text{-}11)$$

In practice, Condition 2, is a requirement that $\psi(t)$ has a zero mean value, i.e., no DC bias, or equivalently that $\hat{\psi}(0) = 0$ or $\int \psi(t)\, dt = 0$.

2.3.2.3 Generating the Wavelet Family

We generate a doubly-indexed family of wavelets from ψ using dilation and translation:

$$\psi_{a,b}(t) \quad = \quad |a|^{-1/2} \psi\left(\frac{t-b}{a}\right) \qquad \text{where } a, b \in \Re,\ a > 0 \qquad (2\text{-}12)$$

The factor $|a|^{-1/2}$ provides the energy scaling factor and ensures that $\|\psi_{a,b}\| = \|\psi\|$ for all a. For the following discussion, we shall assume that $\|\psi\| = 1$. We can represent each wavelet in the family by allowing only discrete values for the dilation and translation parameters:

$$\psi_{m,n}(t) \quad = \quad a_0^{-m/2} \psi\left(a_0^{-m} t - nb_0\right) \qquad (2\text{-}13)$$

The change of parameters corresponds to the choices:

$$a = a_0^m \quad \text{and} \tag{2-14}$$

$$b = nb_0 a_0^m \tag{2-15}$$

indicating that the translation parameter b depends on the chosen dilation.

2.3.2.4 A comparison with the Windowed Fourier Transform

The Wavelet Transform provides a similar time-frequency description as the windowed Fourier Transform, with some important differences. The wavelet transform formulae analogous to Eqs. 2-2 and 2-3 are:

$$\left(T^{wav} f\right)(a,b) = |a|^{-1/2} \int f(t)\psi\left(\tfrac{t-b}{a}\right) dt \tag{2-16}$$

and

$$T_{m,n}^{wav}(f) = a_0^{-m/2} \int f(t)\psi\left(a_0^{-m} t - nb_0\right) \tag{2-17}$$

In both cases, we assume that ψ satisfies:

$$\int \psi(t)\, dt = 0 \tag{2-18}$$

Eq. 2-17 is obtained from 2-16 by restricting a, b to only discrete values: $a = a_0^m$, $b = nb_0 a_0^m$ in this case, with m, n ranging over Z, and $a_0 > 1$, $b_0 > 0$ fixed. One similarity between the wavelet and Windowed Fourier Transforms is clear: both Eqs. 2-2 and 2-16 take the inner products of f with a family of functions indexed by two parameters:

$$g_{\omega,t_0}(t) = e^{i\omega t} g(t - t_0) \qquad \text{(Fourier)} \tag{2-19}$$

$$\psi_{a,t_0}(t) = |a|^{-1/2} \psi\left(\tfrac{t-t_0}{a}\right) \qquad \text{(Wavelet)} \tag{2-20}$$

A typical choice for ψ is $\psi(t) = \left(1 - t^2\right) \cdot \exp\left(-t^2/2\right)$, the second derivative of the Gaussian, sometimes called the Mexican hat function because it resembles a cross section of a Mexican hat. The Mexican hat function is well localised in both time and frequency, and satisfies Eq. 2-18. As a changes, the wavelets $\psi_{a,0}(t) = |a|^{-1/2}\psi(t/a)$ cover different frequency ranges (large values of the scaling parameter $|a|$ correspond to low frequency, or large scale $\psi_{a,0}$; small values of $|a|$ correspond to high frequencies or very fine scale $\psi_{a,0}$). Changing the parameter t_0 allows us to move the time localisation centre: each $\psi_{a,t_0}(t)$ is localised around $t = t_0$. It follows that Eq. 2-2 and Eq. 2-16 provide a time-frequency description of f. The difference between the Wavelet and Windowed Fourier Transforms lies in the shape of the mother wavelets g_{ω,t_0} and ψ_{a,t_0}. The functions g_{ω,t_0} all consist of the same envelope function g, translated to the proper time location, and "filled-in" with the higher frequency oscillations. All the g_{ω,t_0}, regardless of the value of ω, have the same width. In contrast, the ψ_{a,t_0} have time-widths adapted to their frequency: high frequency ψ_{a,t_0} are very narrow, while low frequency ψ_{a,t_0} are much broader. As a result, the wavelet transform is better able than the windowed Fourier transform to "zoom in" on very short-lived high frequency phenomena, such as transients in signals.

The Wavelet Transform: $\psi_{m,n}$ is localised around $a_0^m nb_0$ in time. The Windowed Fourier Transform: $g_{m,n}$ is localised around nt_0 in time, and around $m\omega_0$ in frequency. For a more detailed comparison between the Fourier and Wavelet transforms, see [Dau90, Dau92].

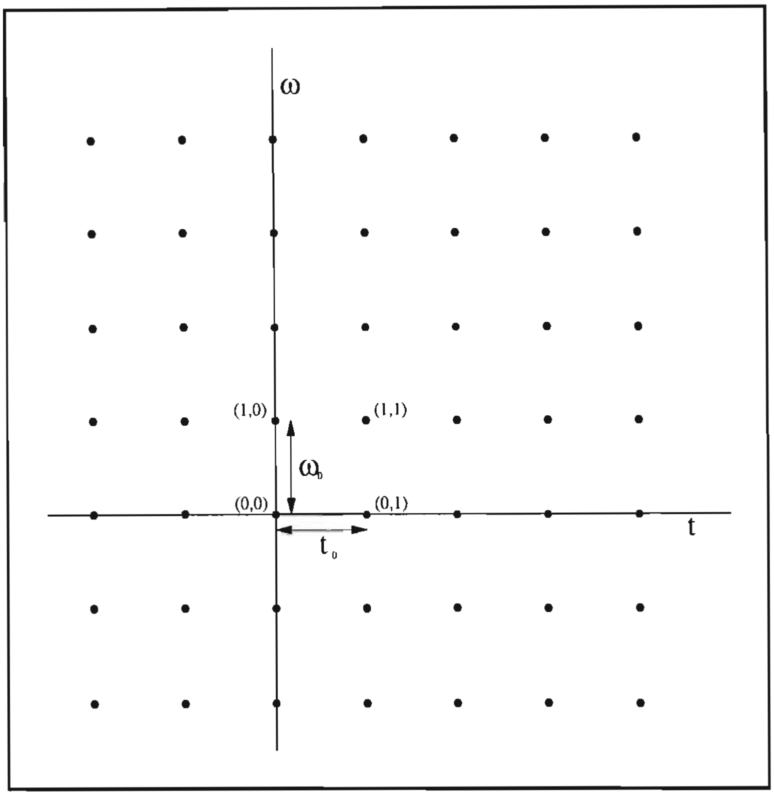


Figure 2-4: The lattice of time-frequency localisation for the Windowed Fourier Transform [Dau92].

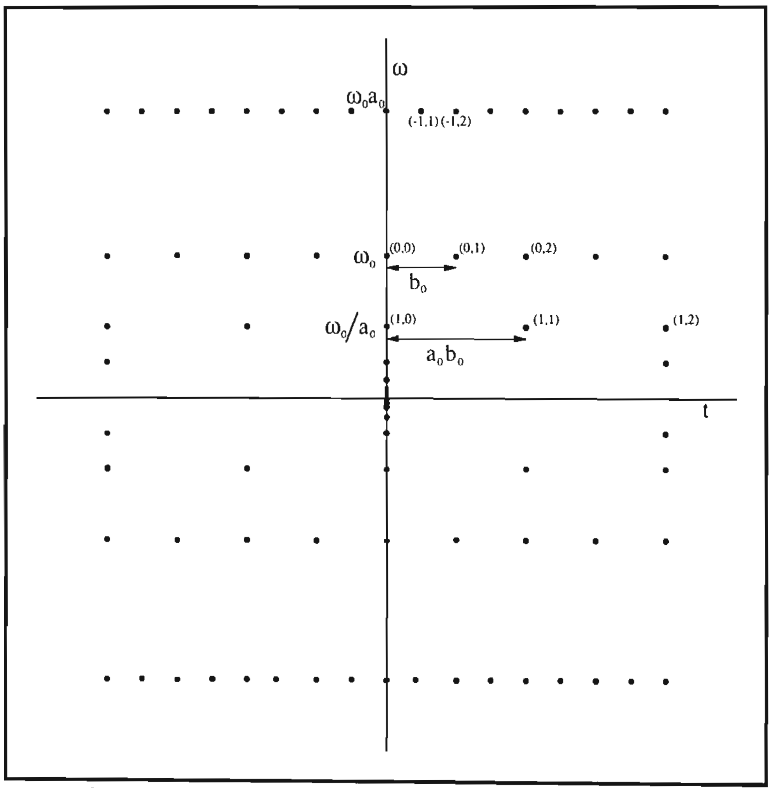


Figure 2-5: The lattice of time-frequency localisation for the Wavelet Transform [Dau92].

2.3.2.5 Orthogonal Wavelet Bases

For some very special choices of ψ and $a_0, b_0 \in Z$, the $\psi_{m,n}$ constitute an orthonormal basis for $L^2(\Re)$. In particular, if we choose $a_0 = 2$ and $b_0 = 1$, then there exists a ψ with good time-frequency localisation properties, such that the functions:

$$\psi_{m,n}(t) \quad = \quad 2^{-m/2}\ \psi\left(2^{-m}t - n\right) \qquad (2\text{-}21)$$

constitute an orthonormal basis for $L^2(\Re)$. The question of when one can obtain orthonormal bases of wavelets using dilation factors other than 2 is also of interest. Auscher studied the case of rational dilation factors a in the region $1 < a < 2$ [Aus92]. Meyer [Mey90] has considered orthonormal bases obtained using integer dilation factors $a > 2$; such bases require two or more sets of wavelets.

2.3.2.6 The Fast Wavelet Transform

The localisation properties of wavelets make them extremely useful for numerical analysis of systems with singular behaviour. In many cases, not only are fewer basis functions required with wavelets than with such traditional bases as Fourier series, but such annoying anomalies as the "Gibb's phenomenon" are minimised. As wavelet theory has advanced, numerical algorithms have simultaneously been developed which exploit these advantages. In particular, Daubechies's compactly supported orthonormal wavelets can be used to develop a Fast Wavelet Transform which appears to be superior to the Fast Fourier Transform for many purposes. The *Fast Wavelet Transform* (FWT) was first proposed by Mallat in his original paper on multi-resolution analysis [Mal89b] using truncated versions of infinitely supported wavelets. A numerical algorithm using the compactly supported wavelets of [Dau88], thereby avoiding the error due to truncation, was subsequently implemented by Beylkin, Coifman and Rokhlin [Bey91].

The FWT is starting to be widely adopted in various applications and is being accepted as a viable alternative to the popular Fast Fourier Transform; see Chapter 1 for references. The FWT is fast, very efficient, produces non-redundant coefficients, and allows users to choose from a large variety of basis functions.

Relative to a mother wavelet ψ, the *Wavelet Transform* (WT) on $L^2(\Re)$ is defined by:

$$\left(W_\psi f\right)_{m,n} \quad = \quad \left\langle f, \psi_{m,n} \right\rangle \qquad (2\text{-}22)$$

for $m,n \in \mathbb{Z}$. If $\psi_{m,n}$ represents a family of orthogonal wavelets, the wavelet coefficients are non-redundant and complete [Chu92]. The *Inverse Wavelet Transform* can therefore reconstruct the signal from the coefficients:

$$f(t) \quad = \quad \sum_m \sum_n \langle f, \psi_{m,n} \rangle \tag{2-23}$$

The two properties which are the hallmarks of the Wavelet expansions are *localisation in time*, and *scaling*. The Fast Wavelet Transform [Str89, Chu92] exhibits many favourable characteristics:

- Orthogonal basis functions are the foundation of the FWT and these make for elegant and non-redundant decomposition.
- Spatial localisation is a by-product as the fundamental building block of the transform is a family of basis functions having finite-support.
- The FWT is an extremely efficient algorithm having an order of complexity equal to the number of samples in the signal $f(t)$, that is, $O(n)$. The Fourier Transform has complexity of $O(n^2)$ and the optimised Fast Fourier Transform has complexity $O(n \log_2 n)$.

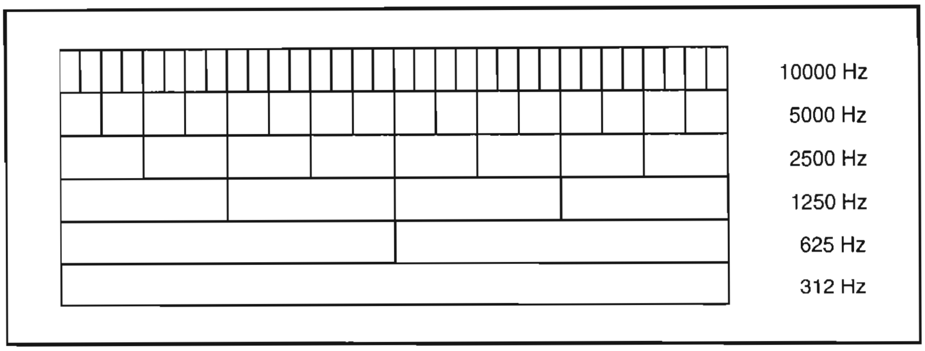


Figure 2-6: Spatial dependence of orthogonal wavelets.
(assuming f_s = 20 kHz)

Mallat's elegant $O(n)$ *tree algorithm* or *pyramid algorithm* performs fast decomposition and reconstruction of signals. It can be thought of as being to the Wavelet Transform what the Fast Fourier Transform is to the Fourier Transform. The algorithm is documented in [Str89]. From the frequency scale, we observe that the wavelet transform is in fact an *octave-band filter*.

The Haar wavelet has been known since 1910. It is orthogonal to its own dilations and translations, but is not continuous. Daubechies [Dau88a] showed that, apart from the Haar wavelet, there exist no other compactly supported wavelet bases in which ϕ is either symmetric or anti-symmetric around any axis. Note that the scaling function for the Haar wavelet is invariant under the coefficients $c_k = \{1 \quad 1\}$. Two other well-studied examples are the cubic B-spline and D$_4$ [Dau88a, the name D$_4$ being proposed in Str89] which have the coefficients $\frac{1}{8}\{1 \quad 4 \quad 6 \quad 4 \quad 1\}$ and $\frac{1}{4}\{(1+\sqrt{3}) \quad (3+\sqrt{3}) \quad (3-\sqrt{3}) \quad (1-\sqrt{3})\}$ respectively. The Haar wavelet is obviously not continuous, whereas the cubic B-spline and D$_4$ wavelets are. The wavelet D$_4$ is not as smooth as it looks, in fact; out of the three, the only one differentiable is the cubic B-spline. Clearly, being continuous or differentiable is not a pre-condition for admissibility. Each has a different support on *t*.

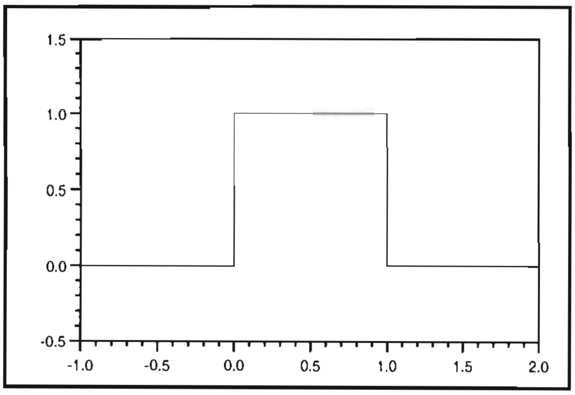


Figure 2-7: Haar scaling function.

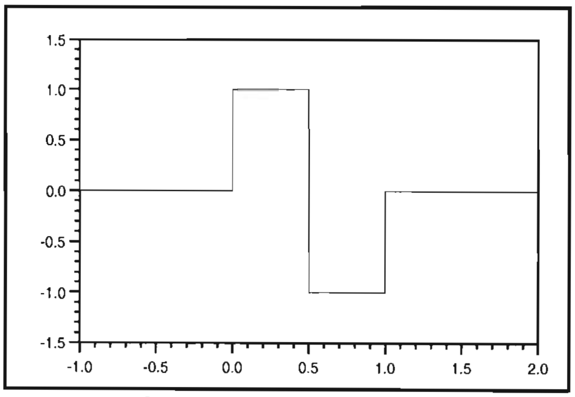


Figure 2-8: Haar wavelet.

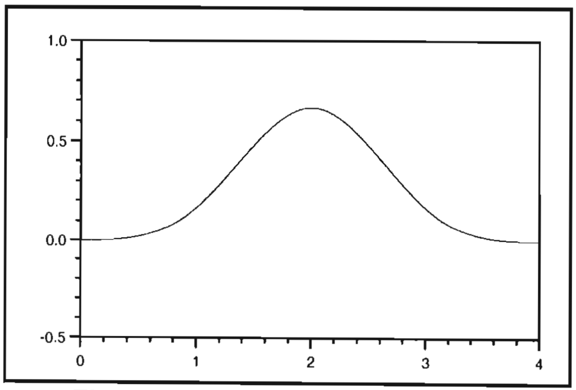


Figure 2-9: Spline scaling function.

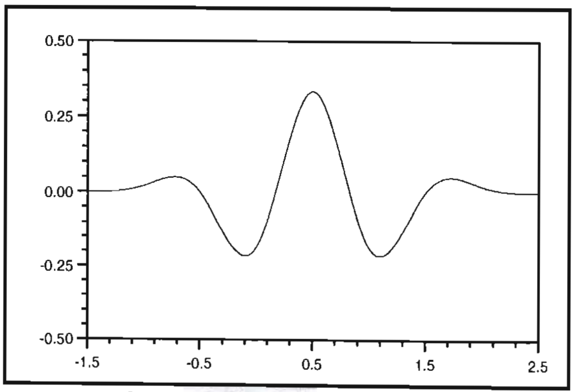


Figure 2-10: Spline wavelet.

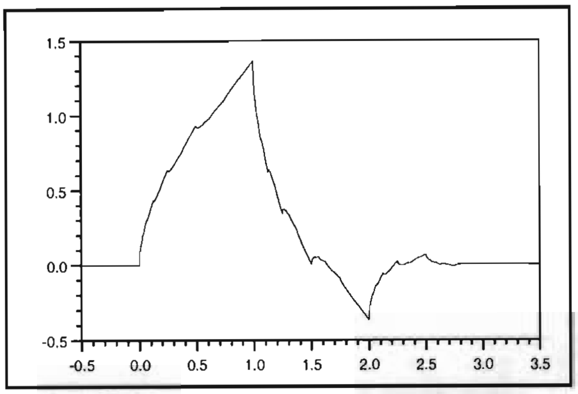


Figure 2-11: Daubechies scaling function.

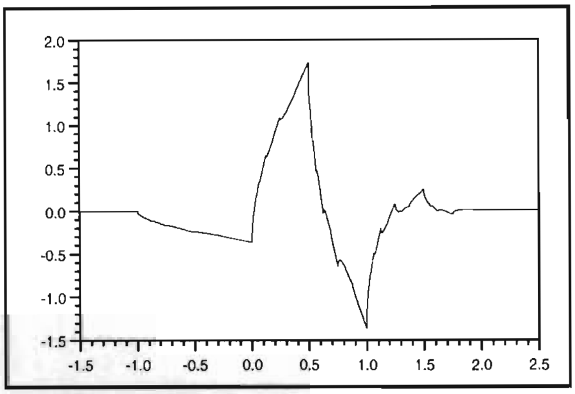


Figure 2-12: Daubechies wavelet D_4.

The Haar wavelet is the simplest and therefore probably the most well-known and understood of all. Starting with a box scaling function ϕ, and the coefficients $c_k = \{1 \ \ 1\}$, we iterate Eq. 2-9 to render the orthonormal *Haar mother wavelet*:

$$\begin{aligned} \psi(t) &= \phi(2t) - \phi(2t-1) \\ &= \begin{cases} 1 & \text{if } 0 \le t < \tfrac{1}{2} \\ -1 & \text{if } \tfrac{1}{2} \le t < 1 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \qquad (2\text{-}24)$$

Figure 2-13 shows translations and dilations of the Haar wavelet. The pattern is obvious. An important characteristic of these wavelets is their mutual orthogonality. This set can be expanded to an arbitrarily large mutually orthogonal basis set. To retain their mutual orthogonality, each wavelet must be fixed spatially.

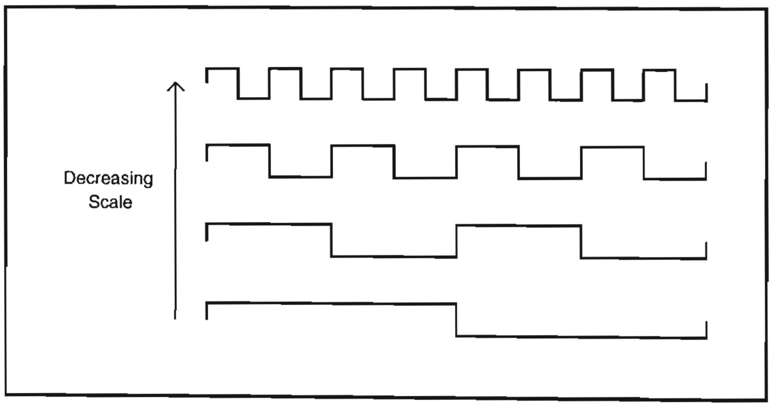


Figure 2-13: Haar wavelet family subset.

2.3.2.7 Why the FWT is unsuitable for Frequency Analysis

Recall Figure 2-6. This diagram expresses the hierarchical tree structure of the Wavelet Transform. Each block in the diagram identifies a wavelet in time-frequency space. These wavelets' spatial localisation must be strictly adhered to in order to maintain mutual orthogonality. In addition, at each level in the pyramid, the scale of the wavelets decreases by a factor of 2.

We propose that the FWT is unsuitable for accurate frequency analysis for the following two reasons:

Spatial Localisation: By definition, the coefficients of the FWT are spatially fixed; i.e. any spatial translation would destroy their mutual orthogonality. As a result, the dot product $\left\langle f, \psi_{a,t_0} \right\rangle$ is dependent on the phase of f.

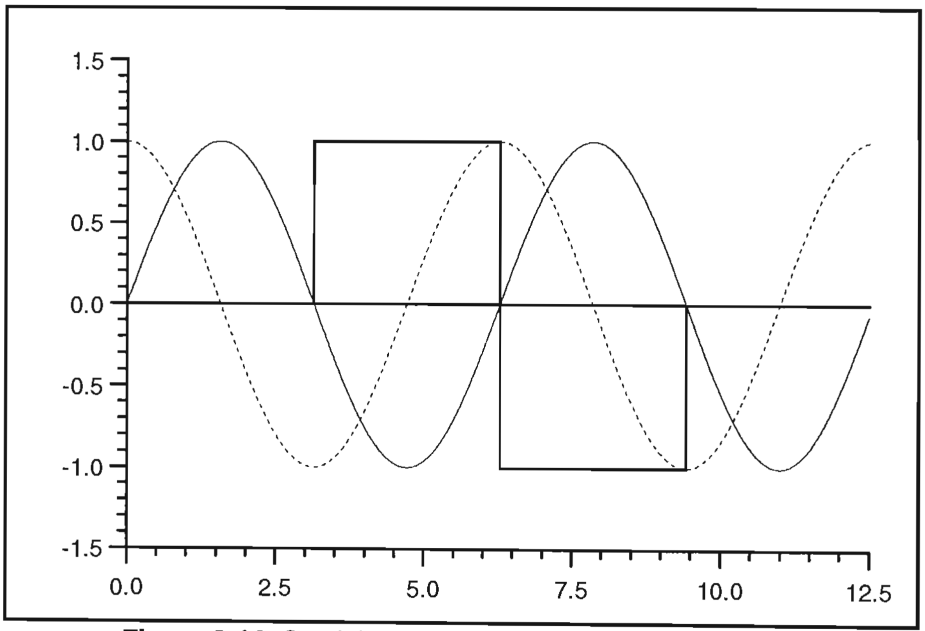


Figure 2-14: Spatial dependence of orthogonal wavelets.

Consider two sinusoidals having a signal length of 16 and magnitude 100 and a phase difference of $\pi/2$ radians, i.e. two signals equivalent to sine and cosine waves. Both have the same frequency and constitute exactly 2 cycles. Consider the following tables:

-70	29	70	-29	-70	29	70	-29	a = 1
-70		70		-70		70		a = 2
241				241				a = 4
0								a = 8

Table 2-1: Haar wavelet coefficients of sine wave.
(Note: All decimal places truncated.)

29	70	-29	-70	29	70	-29	-70	a = 1
170		-170		170		-170		a = 2
100				100				a = 4
0								a = 8

Table 2-2: Haar wavelet coefficients of cosine wave.
(Note: All decimal places truncated.)

Table 2-1 shows the wavelet coefficients of the sine wave and by examination of the rows, it appears obvious that the frequency of the wave has scale $a = 4$. However, a translation of the wave by $\pi/2$ radians (a quarter cycle) results in the most significant coefficients now indicating that the wave has scale $a = 2$. Evidently, the frequency of the wave has changed, but clearly it has not. This simple example shows the 'translation dilemma' of the WT. However, we must stress that there is nothing 'wrong' with the coefficients, but they unfortunately are not immediately representative of the signal as far as spectra analysis is concerned. It is this discrepancy, due to the spatial localisation, which prohibits us from interpreting the coefficients at different scales of the WT for spectral analysis. Therefore, as orthogonal wavelets do not provide consistent coefficients for any given wave, irrespective of phase, the elegant tree algorithm of Mallat unfortunately cannot be applied with any notable degree of consistency.

Sparse Frequency Space: Consider $\frac{1}{10}$s worth of sampling with f_s = 22 kHz. Because the frequency scale of the FWT increases like 2^a, the frequency space consists of the frequencies 11 kHz, 5.5 kHz, 2.75 kHz, 1.37 kHz, 687 Hz, 343 Hz, 171 Hz, 86 Hz, and so on. We refer to this small number of frequency bands as a *sparse frequency space*. Compare this with the Fourier Transform whose frequency space would consist of all frequencies from 0 kHz to 11 kHz in linear increments of only 10 Hz. This *dense frequency space* would render a more accurate representation of the signal in the frequency domain, and its linearity would not exhibit the severe decay of the higher frequencies which the WT possesses.

2.4 The Dominant Scale Transform

The problems associated with Fourier Analysis and the spectra inconsistencies introduced by the Wavelet Transform prompted the author into researching an algorithmic approach to time-frequency decomposition. The result was the $O(n)$ integer-based *Dominant Scale Algorithm* (DSA) which decomposes a signal $f(t)$ into its approximate time-frequency space $F_\psi(t,\omega)$ or, perhaps more accurately, its time-scale space $F_\psi(t,a)$. The coefficients in this space form the *Dominant Scale Transform* (DST). This time-frequency space allows us to look at the spectral contribution of a specific portion of the signal without concern about any samples outside the small interval in question. In addition, the 'frequency-varying windowing' or zooming, allows us to spatially localise transients without concern about spectral bleeding.

2.4.1 The Dominant Scale Algorithm

The two obstacles concerning the Wavelet Transform are nullified by making two changes to the basis functions, each solving one of the two problems associated with time-frequency decomposition.

- The simplest way to remove the spatial dependency is to consider all the *integral shifts* of ψ, namely: $\psi(t-k)$, $k \in \mathrm{Z}$. By relaxing the strict imposition of spatial positioning of the orthogonal wavelets, we obtain better spatial localisation.

- Secondly, we change the wavelength of the wavelet at scale a to $2a$ from the conventional 2^a. This increases the number of coefficients which in turn increases our frequency resolution.

Both of these changes increase the number of basis functions and destroy their mutual orthogonality, but in return we obtain extremely well localised time-frequency coefficients as well as a dense frequency space. Although the Haar wavelet is considered to exhibit bad time-frequency localisation [Dau92], we adopt the simple basis function as a matter of efficiency. For all t, the Haar wavelet $\psi(t) \in \{-1, 0, +1\}$. As a consequence, the calculation of $\langle f,\ \psi_{a,b}\rangle$ involves mainly additions and subtractions. If the function $f(t)$ is integer-based, which is most common for ADC output, the dot product too will be entirely integer-based. Note that should $f(t) \in \Re$, the arithmetic simply becomes real-valued and no complications are introduced. There are several more factors affecting our choice of ψ, most of which are described later in the chapter.

Whether or not changing the mother wavelet shape would notably alter the decomposition characteristics was not examined in much detail, although preliminary tests show that such an alteration is, in fact, unlikely to occur. Moreover, the computation time was increased significantly by the introduction of floating-point calculations. We therefore intentionally restrict ourselves to the Haar wavelet. We can now define the concept of a dominant scale:

Definition: Given a set of samples f, a wavelet basis ψ, and a scale range $\{MinA..MaxA\}$ then if:

$$\alpha(t_i) = \left\{ a \mid \left\langle f, \psi_{a,\, t_i} \right\rangle = \max_{b \,\in\, \{MinA..\,MaxA\}} \left\langle f, \psi_{b,\, t_i} \right\rangle \right\}$$

then the *dominant scale* of f is:

$$\chi(t_i) = \begin{cases} \alpha(t_i) & \text{if } \left\langle f, \psi_{\alpha(t_i),\, t_i} \right\rangle \geq \left\langle f, \psi_{\alpha(t_i),\, t_{i\pm 1}} \right\rangle \\ 0 & \text{otherwise} \end{cases}$$

In other words, to obtain the *Dominant Scale Function* $\chi(t_i)$, any entry in the *Maximum Scale Function* $\alpha(t_i)$ is ignored if it does not correspond to a local maximum in the *Maximum Energy Function* $\left\langle f, \psi_{\alpha(t_i),\, t_i} \right\rangle$. The value $\alpha(t_i)$ is the scale containing the most energy at t_i and thereby being the most probable scale at that time instant. In short, the first pass can almost be described as a best-fit process, where the signal is superimposed with a single-cycle wavelet, and the energy of their dot-product calculated. The wavelet with the greatest energy at that time t_i is declared the 'winner'. The calculations are performed for $t_0, t_1, t_2, \ldots$ This defines two functions, namely the Maximum Scale Function and the Maximum Energy Function. The locating of maxima in the Maximum Energy Function can be thought of as the localising of that frequency coefficient's zero-crossing; in other words, the dot product will reach a maximum at that component's zero-crossing. Admittedly, this is somewhat of an ad hoc approach. The idea of retaining only the dominant scale was previously documented in [Mal92] which describes the detection of singularities in one-dimensional and two-dimensional signals.

Although the DST involves a significant amount of over-sampling, most of the resulting coefficients are eliminated by the *max* function, with only those coefficients associated with local maxima in the energy function being retained. Unfortunately, the above definition relies partly on the fact that each frequency component in the signal is of roughly equal magnitude. In practice, however, the absence of this constraint is not necessarily a predictor of failure. For example, the algorithm was applied to speech signals in which harmonics having much lower magnitudes than the fundamental were present and, as we shall see in the following chapter, the DST performed very well indeed. The definition of the dominant scales can be implemented very simply by the algorithm below to generate the Dominant Scale Transform time-frequency space.

```
Dominant Scale Algorithm:
For each time $t$ of a signal $f$
    Calc the scale $a$ having max energy and set:
        MaxScale    [ $t$ ]  =  $a$
        MaxEnergy   [ $t$ ]  =  $\langle f, \psi_{a,t} \rangle$
For each entry $t$ of MaxScale
    If LocalMaximum (MaxEnergy [ $t$ ])
            DomScale [ $t$ ] = MaxScale [ $t$ ]
        Else
            DomScale [ $t$ ] = 0
End.
```

From now, we shall refer to the MaxScale function as the "Scale Function", the MaxEnergy function as the "Energy Function", and DomScale function as the "Time-Freq Function". In practice, the second pass is easily incorporated into the first. As an example, consider a sine wave, and its associated energy function.

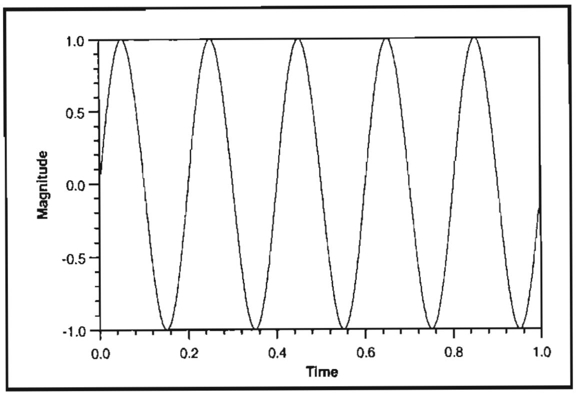


Figure 2-15: Sine wave.

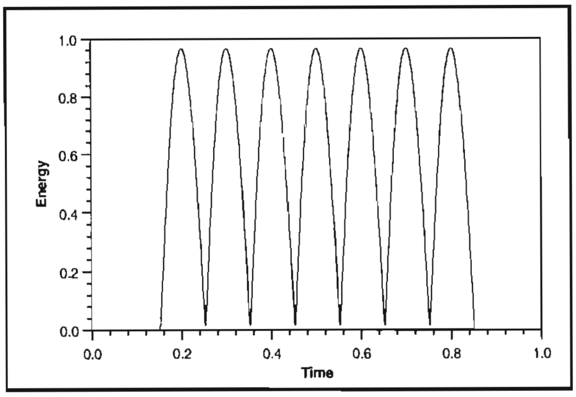


Figure 2-16: The energy function of the sine wave in Figure 2-15.

Note that both the original signal $f(t)$ and the energy function $\left|\langle f, \psi_{a,t} \rangle\right|$ have been normalised for display purposes. The maxima in the energy function correspond to zero-crossings in Figure 2-15. The $|.|$ operator ensures that both positive and negative maxima are isolated. The sign of the dot product describes the gradient at the crossing point. If the sign of the dot product is retained, we could choose to localise only crossing points having a positive gradient, or vice versa.

2.4.1.1 DSA Order of Complexity

Assume the signal $f(t)$ has length *N*. It is easy to verify that the process has *O*(*N*) complexity.

Proof: For every t_i where $i \in \{1..N\}$, the dot product for each scale $\alpha \in \{MinA..MaxA\}$ is calculated. This has complexity *O*(*MaxA*-*MinA*+*1*), and as *MaxA* and *MinA* are pre-defined constants, we simply write it as *O*(*k*) where *k* is a constant independent of *N*. This process is repeated at each t_i and the entire process therefore has *O*(*kN*) complexity. ❑

2.4.1.2 DST Frequency Range

Determining the scale range $[MinA..MaxA]$ and hence the frequency range is quite straightforward. Let the scale *MinA* indicate the finest scale of interest and the scale *MaxA* the coarsest scale of interest; then we are interested in all scales in the range *MinA*, *MinA*+1,..., *MaxA* or $[MinA..MaxA]$. Since *MinA* represents the most compact wavelet, or frequency ceiling, we set $MinA = 1$. This sets the frequency ceiling at the Nyquist frequency limit $f_s/2$. The actual upper frequency limit is restricted by anti-aliasing filters, and is typically about 80% of the Nyquist frequency [Ran87, Lan88]. Ideally, f_S is increased accordingly. To fix *MaxA*, we choose a frequency floor, and set *MaxA* according to this maximum wavelength.

2.4.2 DST Properties

We now turn our attention to some properties of the DST and in doing so, we re-visit the examples which proved troublesome for the Fourier Transform in Chapter 1.

2.4.2.1 Spike

A pure sine wave was generated and a spike having positive magnitude introduced at a local minimum. The three plots in Figure 2-17 show the DST of the following functions:

1. A pure sine wave.
2. A pure sine wave with a positive spike introduced at a local minimum with magnitude of the sine wave.
3. A pure sine wave with a positive spike introduced at a local minimum with 10 times the magnitude of the sine wave.

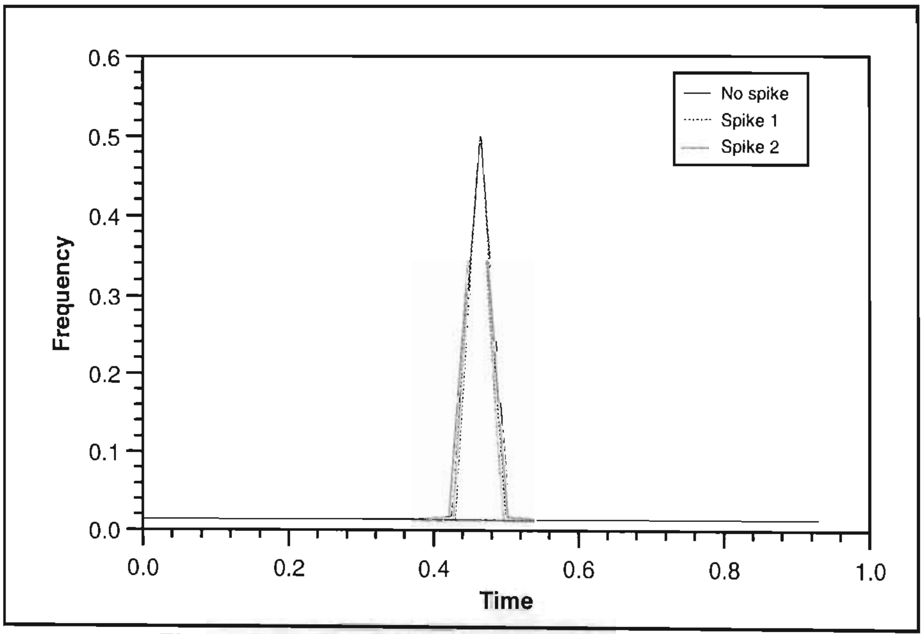


Figure 2-17: The DST of spike affected sine wave.

Both spikes have been excellently localised. If the length of the signal were to increase, the shape of the graph would remain constant and the lines of the low frequency portions would simply extend as the signal lengthened. This is in contrast to the Fourier Transform which would adjust the Fourier coefficients if either the length of the signal changed or the magnitude of the spike was altered. Neither of these two seem to significantly affect the DST coefficients. Moreover, irrespective of the magnitude of the spike (or noise) introduced at, say, t_0 due to the localisation properties, it is impossible for the irregularity to affect any coefficients outside the small interval $\left[t_0 - MaxA..t_0 + MaxA\right]$.

2.4.2.2 Non-stationary signals

Non-stationary signals, in which frequency changes occur with time, were shown to be troublesome for the Fourier Transform. Moreover, applications where frequency varying signals are used are probably the most common problems for the Fourier Transform; hence the enormous amount of research into the Windowed Fourier Transform. As we shall discuss later in the text, complex wave patterns effect some form of trade-off when choosing the window length. Some examples are: speech, music, and bird whistle recognition. The DST of the signal appearing in Figure 1-5 in Chapter 1, is shown below in Figure 2-18. Note the increasing step size as the frequency increases. This is characteristic of the $\frac{1}{2a}$ scaling factor.

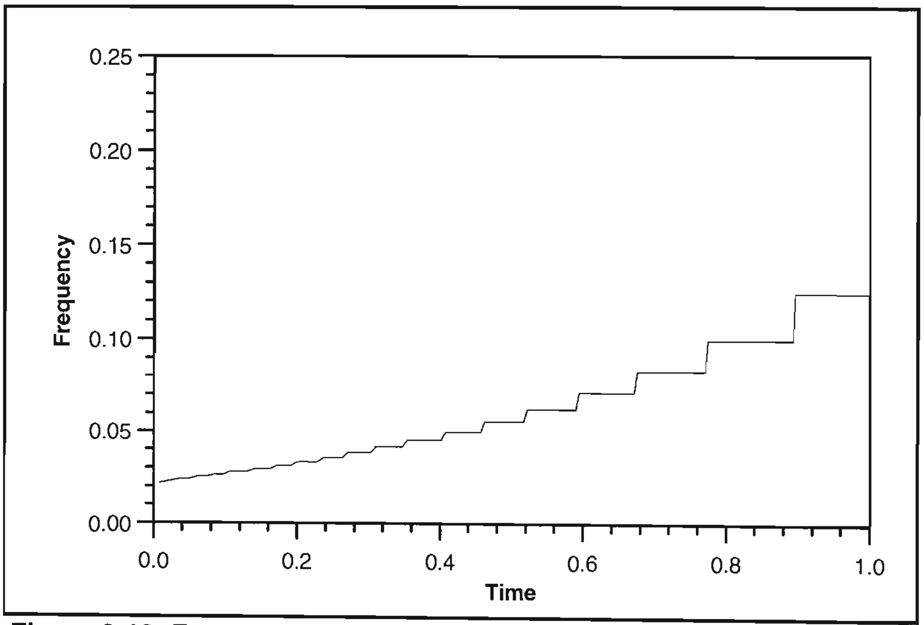


Figure 2-18: Frequency response of a sine wave with increasing frequency.

2.4.2.3 Phase Changes

For each frequency component, the wavelet dot product contains more energy at the zero-crossing point than at any other time. Therefore, each time a dominant scale is isolated during the second pass, the *phase* of the component is automatically a by-product.

Explanation of the time-frequency graphs: The first in a whole series of time-frequency graphs appears in Figure 2-19. You shall see several 'small black dots' in the time-frequency space. These represent the non-zero coefficients of the DST. These coefficient marks do not convey any information about their magnitude, or alternatively, the energy corresponding to that particular scale at that particular time. The author tried several 'nicer' display types. One of these, and probably the most obvious, was a surface plot of the time-frequency space. For example, plotting the time-frequency space of a signal having constant frequency, we would expect to see a ridge running along the entire *t*. However, as the DST consists essentially of sparsely located coefficients, the surface plot showed isolated spikes which did not prove to be aesthetically pleasing and although the eye is excellent at recognising trends, the surface plot was simply not an acceptable form of output. Therefore, we have resorted to the 'dots' approach. This technique actually performs well when dealing with speech (Chapter 3), as one can clearly see the various frequencies present as well as the trends they follow.

Returning to the example in Figure 1-5 and the corresponding FT frequency response in Figure 1-6, we illustrate its DST coefficients in Figure 2-19. Note the accurate spatial localisation of the frequency transients, and the decrease in Δt between non-zero coefficients as the frequency increases, or alternatively, as the wavelength decreases.

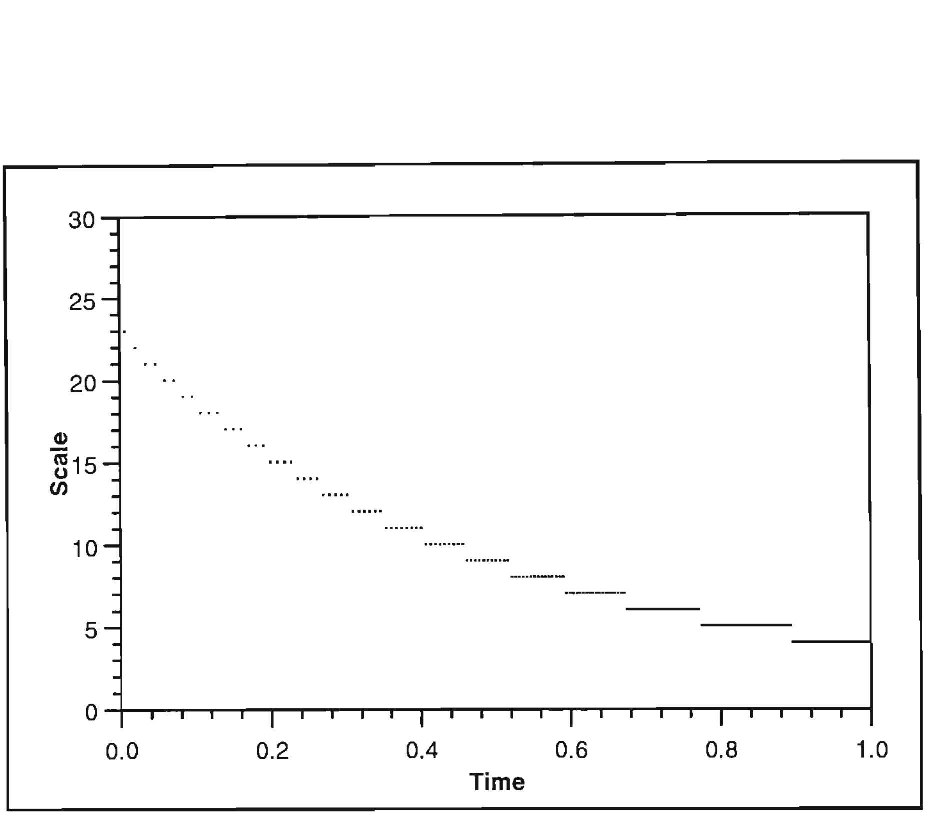


Figure 2-19: Time-frequency space of a sine wave having increasing frequency.

Our next set of examples consists of sinusoidal signals which undergo a phase change. Our first example shows a sine wave which undergoes a π/2 radian phase change, while in the second example a sine wave undergoes a phase change of π radians.

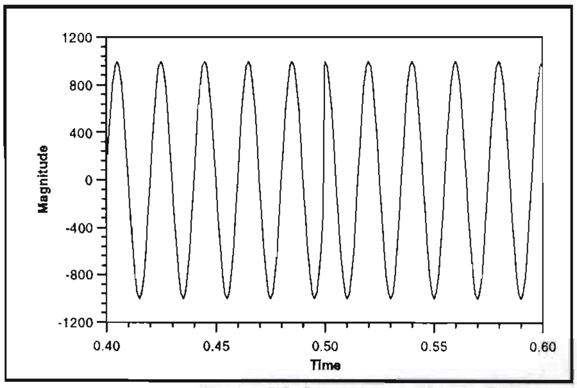


Figure 2-20: Sinusoidal wave undergoing a π/2 radian phase change.

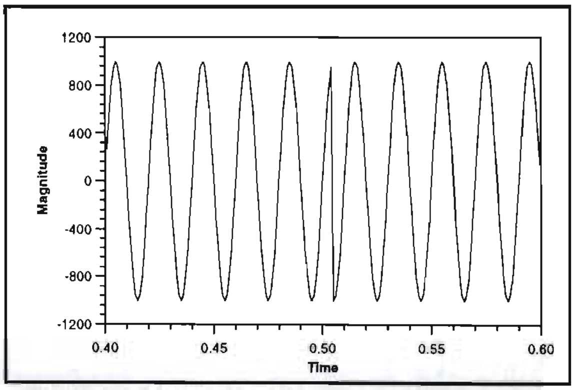


Figure 2-21: Sinusoidal wave undergoing a π radian phase change.

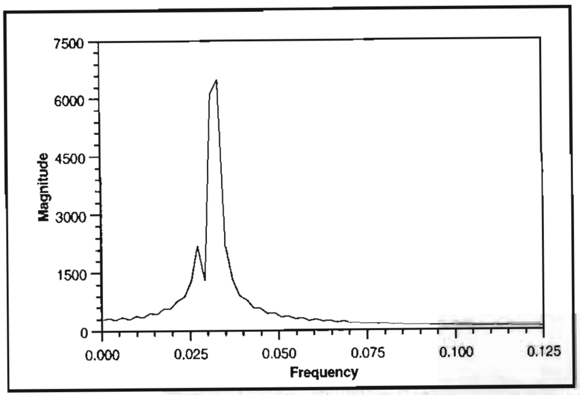


Figure 2-22: Fourier Transform of sinusoidal wave having a $\pi/2$ radian phase change.

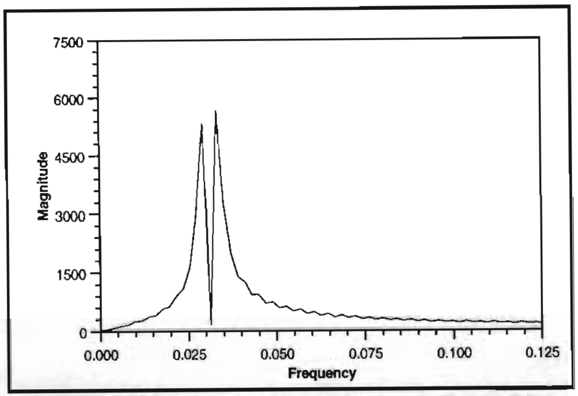


Figure 2-23: Fourier Transform of sinusoidal wave having a π radian phase change.

The Fourier Transform behaves well when subjected to a $\pi/2$ radian phase change. However, a phase change of π radians causes two spikes on the Fourier Transform to appear. The phase of either signal is practically impossible to extract from the Fourier coefficients (barring using the Inverse Fourier Transform!).

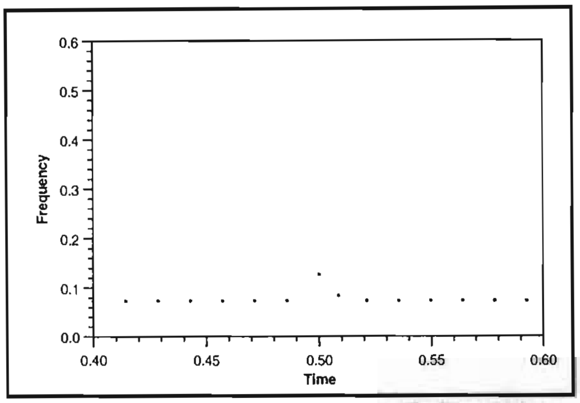


Figure 2-24: DST coefficients of sinusoidal wave having a $\pi/2$ radian phase change.

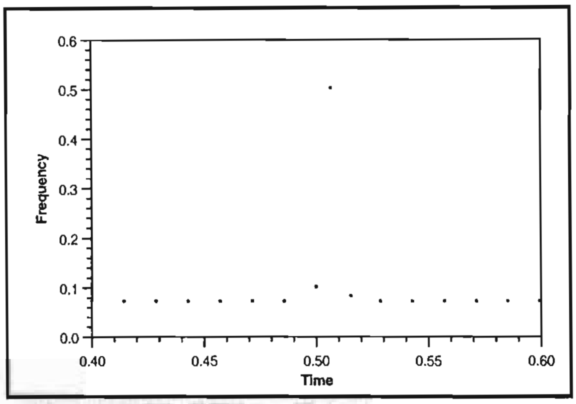


Figure 2-25: DST coefficients of sinusoidal wave having a π radian phase change.

Figures 2-24 and 2-25 show clearly the advantages of the time-frequency space decomposition provided by the DST. Close examination of Figure 2-24 will indeed show that the signal has undergone phase change at $t = 0.5$. The reader is encouraged to make physical measurements of the graph with a pen and paper to verify the phase change. A $\pi/2$ radian phase change induces a rapid change in magnitude at the Nyquist frequency. This appears in Figure 2-25 as a single coefficient at the Nyquist frequency.

Phase information is extremely valuable in many applications and is of prime importance in communications systems for several reasons:

- In digital communications, the phase of the transmitted signal may be used to encode the binary data sent using techniques such as phase-shift keying in which the data stream toggles the phase of a constant amplitude sinusoid between two values say 0 and 180 degrees to indicate a 0 or a 1. This is called binary phase-shift keying. If 4 phases are used, say ±45 and ±135 degrees, the binary combinations 00, 01, 10, and 11 could be allocated, thus increasing the data throughput with the same signalling rate (quadra-phase shift keying QPSK) [Lee88].

- Phase knowledge is useful in modelling and measuring high frequency devices such as antennas, microwave transistors, transmission lines etc. in which the amplitude and phase of a known signal is compared to that reflected from the device, in order to determine its input and output impedance (AC resistance). This method is known as the S-parameter measurement method.

- In control of electric motors, phase knowledge is critical. The phase of the currents applied to the field windings of an AC motor can alter its speed and torque characteristics. In addition, measurements of the phases of the currents induced by the machine onto its windings can indicate the 'health' of the machine, its load, and its operational stability.

- In many communication channels, the phase characteristics of the channel itself are important to data integrity. For example, one could not expect low error rates if one was using phase-shift modulation of 90 degree spacing and the channel regularly showed rapid changes in phase of, say, 100 degrees. The phase response of a channel is intimately linked with the group delay of a signal sent over it, and as such, group delay equalisers are required to ensure consistent phase response over the operational bandwidth.

2.4.2.4 Signal Clipping

Clipping occurs when the range of the recorded signal exceeds the capabilities of the equipment. The fault may lie in either the hardware or software components of the system. For example, the gain on the pre-amplifier may be set too high, thereby possibly exceeding the allowable input voltage of the ADC. Alternatively, the analogue-to-digital mapping may produce a legal hardware signal beyond the range of the software. For example, the mapping attempts to assign a value of 300 to an 8-bit entity. Clipping can be defined as follows:

$$f(t_i) = \begin{cases} T & \text{if} f(t_i) > T \\ -T & \text{if} f(t_i) < -T \\ f(t_i) & \text{otherwise} \end{cases} \qquad (2\text{-}25)$$

We simulated clipping by generating a pure sine wave and limiting the signal to 90%, 80%, 70%,..., 10% respectively of its original magnitude. An example of a signal clipped at 70% of its maximum magnitude, together with its corresponding Fourier Transform, is shown below.

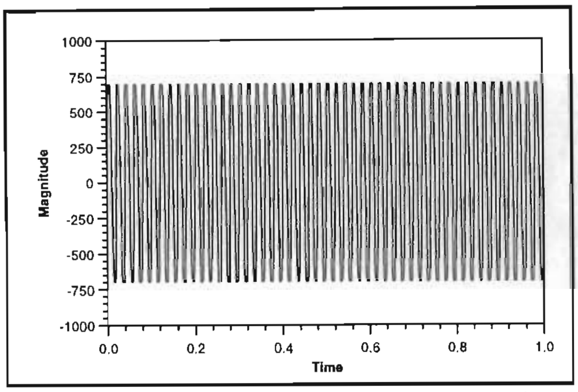


Figure 2-26: A sine wave clipped at 70% of peak amplitude.

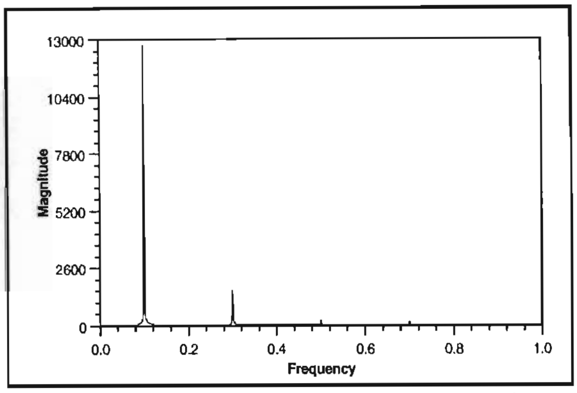


Figure 2-27: The Fourier Transform of the wave in Figure 2-26.

The Fourier Transform correctly detected the fundamental frequency and introduced low magnitude harmonics. Overall, the transform's behaviour was very good. To test the behaviour of the DST when subjected to extreme amounts of clipping, a number of pure sine waves were generated and clipped at various magnitudes. The sine wave frequencies were: 5000 Hz, 4000 Hz, 3000 Hz, 2000 Hz, 1000 Hz, 500 Hz, 200 Hz, 100 Hz, and 50 Hz, and the levels of clipping ranged from 0% to 90% in 10% steps. The results appear in Figure 2-28. The Fourier Transform performed better in this regard when compared with the DST. Although the DST's accuracy was quite satisfactory, clipping should be avoided if at all possible.

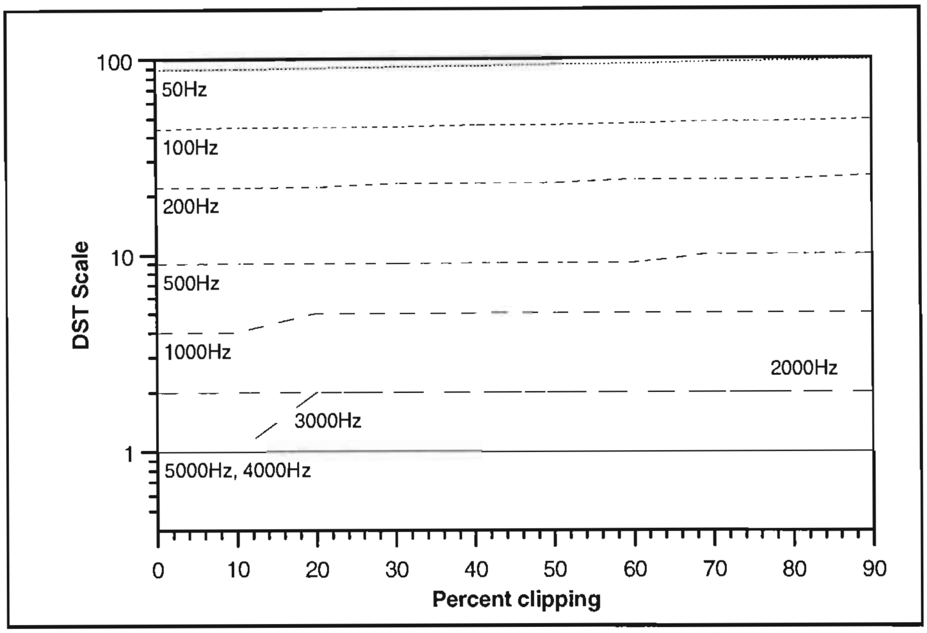


Figure 2-28: Degrading affect on decomposition of clipping.
(Note: Logarithmic scale of Y-axis)

2.4.2.5 DST Frequency Scale

The wavelet dilation factor $\frac{1}{2a}$ results in the DST's coefficients representing a logarithmic frequency scale. This is in contrast to the Fourier Transform which has a linear frequency response. The logarithmic response is remarkably similar to humans' hearing perception; this phenomenon is investigated further in Chapter 3. The octave filtering, with $1/2^a$ dilation factor, of the Wavelet Transform causes a very steep fall-off in the frequency scale. The $\frac{1}{2a}$ wavelet scaling factor of the DST results in a less drastic frequency decay (see Figure 2-29).

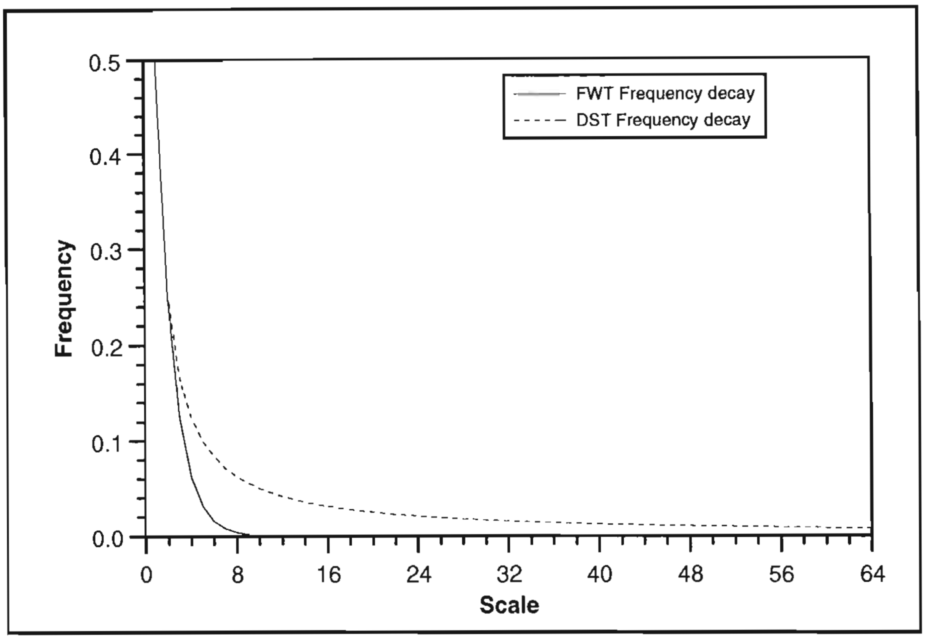


Figure 2-29: Frequency decay rates of the FWT and DST.

2.4.2.6 DST Underscaling

The DST exhibits a certain characteristic, which we call *DST underscaling*, that causes a slight underestimating of the scale, which in turn translates into a slightly higher frequency estimate. Underscaling is a result of the diminishing energy towards the ends of sinusoidal cycles. Fortunately, the error can be measured and corrected. Figure 2-30 shows energy versus scale plots.

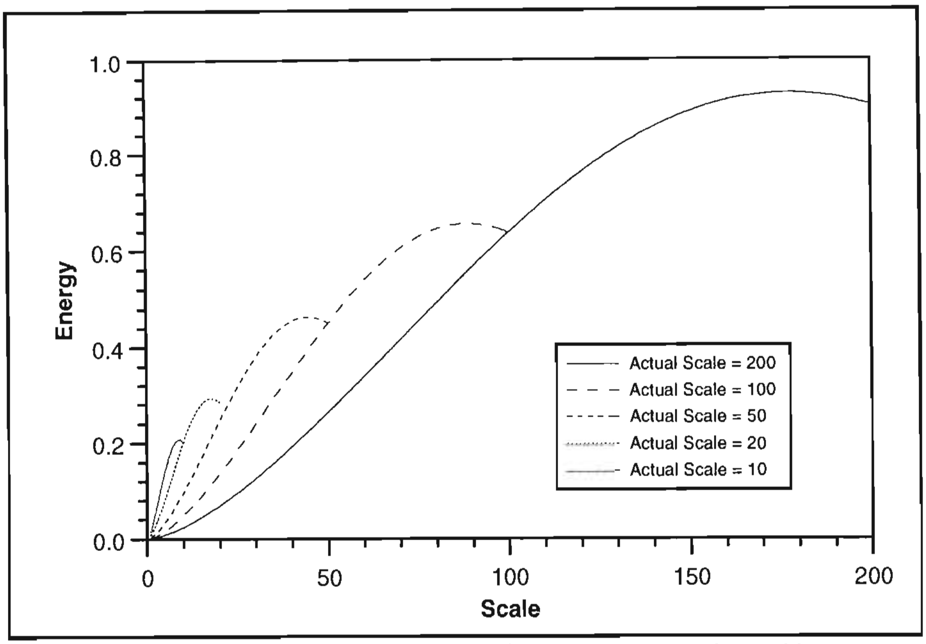


Figure 2-30: Energy functions of expanding wavelets with sine wave of constant wavelength.

The larger (rightmost) curve represents the energy of a sine wave with wavelength 400 with wavelets of varying scale. Ideally, the scale having maximum energy should equal 200; however, it is slightly underestimated. The other curves show similar examples, where the sine waves' wavelength has decreased each time. The amounts by which the DST underestimates the scales are shown in Figure 2-31.

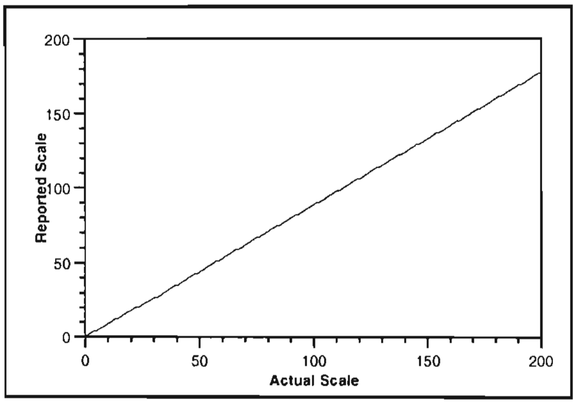


Figure 2-31: DST scale correction.

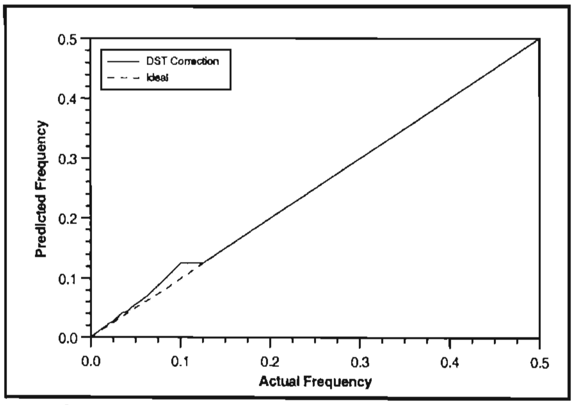


Figure 2-32: DST frequency correction.

Fortunately the scale correction curve approximates a straight line, thereby making the scale correction straightforward. Whereas Figure 2-31 shows the scale correction curve, Figure 2-32 shows the frequency correction curve. The ideal $X = Y$ line is also drawn. Figure 2-32 can be interpreted as follows. The smallest scales are reported correctly and since these contribute to a significant portion of the high frequencies, the DST estimates the ideal above $f_s/8$. We next turn our attention to the lower frequencies. Although the difference between the actual scale and reported scale increases as we enter the larger scales, the difference between adjacent coefficients' frequencies tends to zero (note the asymptotic trend in Figure 2-29). Therefore, although the difference in scale increases towards the larger scales, the difference in frequency decreases, causing the error to decrease in sympathy. Consequently, there exists only a small portion of the frequency range that needs to undergo any meaningful frequency correction.

The largest error occurs at 10% of f_s where the frequency is overestimated by exactly 25%. We re-iterate that DST underscaling can be corrected using either Figure 2-31 or 2-32.

2.5 DSA Implementation Specifics

We showed earlier that the generating of the DST coefficients has $O(n)$ complexity. We now wish to briefly discuss some of the implementation specifics of the DSA as the coding can often make a substantial difference to the efficiency of the algorithm. There are three factors which we feel contribute greatly towards the efficiency of the algorithm. Many of these factors were taken into account during the development of the algorithm. We briefly discuss these below.

2.5.1 Hardware implementation of DSA

The algorithm is perfectly suited for hardware implementation for several reasons:

- Practically all operations are *simple operations*; additions and subtractions are used in preference to multiplications and divisions. Additions and subtractions are more efficient to execute in software and are easier to implement in silicon.

- The algorithm is entirely *integer-based.* This totally eliminates computationally expensive floating-point operations. If the sampled signal is represented in floating-point, the transform can work in this domain if so desired. Hardware implementation of integer-based arithmetic is fairly effortless.

- The algorithm is *highly parallel* in design which can significantly reduce run-time. By cascading the adder units, a significant amount of parallelism can be achieved. Each of the $\langle f, \psi_{a,b} \rangle$ calculations is entirely independent of the other dot products for different a, b. In the extreme case (assuming an unlimited chip size), we could achieve an order of complexity $O(k)$ where $k = MaxA\text{-}MinA+1$, which is independent of N (the number of samples in the signal). The only inter-dependency would be the computing of the *max energy* function. The operations required for this computation can be executed in parallel to a very large degree.

2.5.2 Using a more suitable language to optimise bottleneck code

The code generated by modern-day compilers is very well optimised, however, its efficiency still cannot compete with hand-coded routines. During the implementation of the algorithm, the original Pascal source code was converted to Intel 80x86 assembler code. Obviously, the speed-up obtained depends on the redundancy in the compiler-generated code as well as the skills of the programmer. With the minimal amount of effort, the author was able to obtain a huge increase in speed by coding the entire bottleneck in assembler in only 32 lines!

2.5.3 Optimising bottleneck code

During the optimisation of the bottleneck, we discovered the following two improvements:

- During the calculation of Maximum Scale Function, we can store the dot product of the previous iteration for scale a and then simply add or subtract the next sample outwards from t_0 to obtain the dot product for scale $a+1$. This technique avoids an enormous amount of redundant arithmetic.

- A slightly less influential inefficiency is the calculation of the energy scaling factor $\sqrt{a}$. We are faced with three possible solutions, namely: 1) calculate $\sqrt{a}$ on each iteration, 2) store $\sqrt{a}$ values in a table indexed by a, 3) square the numerator and then simply divide by a. Option 1 is clearly the most inefficient, with preference for options 2 and 3 determined by which of two operations is more efficient, namely a memory access or the squaring of an integer respectively. By using $\sqrt{a}$ in option 2, we introduce floating-point arithmetic which we have strived to eliminate altogether. Note that, if we square both numerator and denominator, the ordering of the values would not change, therefore no square root is required to correct the squaring.

2.5.3.1 Justification for elimination of $\sqrt{a}$

The wavelet family is defined by the dilation, translation, and energy scaling parameters a, b and $|a|^{-1/2}$, respectively, by:

$$\psi_{a,b}(t) \quad = \quad |a|^{-1/2}\psi\left(\tfrac{t-b}{a}\right) \tag{2-26}$$

Although we could store all possible $\sqrt{a}$ in a table, thereby substituting a more efficient memory access instead of a *sqrt* calculation, we try to avoid the calculation of $\sqrt{a}$ because this introduces computationally expensive floating-point arithmetic. We take the standard wavelet coefficient calculation (we assume $a > 0$):

$$c_{a,b} \quad = \quad \frac{\left\langle f, \psi\left(\tfrac{t-b}{a}\right)\right\rangle}{\sqrt{a}} \tag{2-27}$$

and square:

$$c_{a,b}^2 \quad = \quad \left(\frac{\left\langle f, \psi\left(\tfrac{t-b}{a}\right)\right\rangle}{\sqrt{a}}\right)^2 \tag{2-28}$$

$$\quad = \quad \frac{\left\langle f, \psi\left(\tfrac{t-b}{a}\right)\right\rangle^2}{a} \tag{2-29}$$

Since the second pass of the DST concerns itself solely with maxima present in the energy function $\left|\left\langle f, \psi_{a,b}\right\rangle\right|$, squaring of the coefficients will not change the location of these maxima. We can now calculate the coefficients using only integer arithmetic assuming the domain of $f(t)$ to be integer.

2.6 Chapter Summary

The chapter began by briefly introducing the reader to wavelets. Disadvantages of the traditional Fast Wavelet Transform were illustrated. We then described the new Dominant Scale Transform and the associated Dominant Scale Algorithm which decomposes a time-domain signal f into its time-frequency space. The algorithm was designed with both accuracy and efficiency in mind. The latter part of the chapter closely examined several properties of the DST as a spectral analysis tool as well as some of the implementation concerns. The following chapter applies the DST to a real-world problem that has been in existence for decades and which, despite being very-well researched, remains problematic: Speech Recognition.

Chapter 3

Speech Processing and Recognition using An Approximate O(n) Speech Spectral Analyser: The DST

3.1 Introduction

In this chapter, we begin by briefly introducing two existing methods which have been used very successfully in speech recognition, namely Hidden Markov Models and Cepstrum Analysis, techniques which have proved to be extremely effective and useful in speech processing and other applications. We then investigate the use of the Dominant Scale Transform (DST) as a tool in speech recognition. The research was conducted on a small vocabulary and using three speakers, two male and one female. A model, based on various characteristics, of each word was defined which we used to classify any future incoming signal. Our model attempts to stress the characteristics which are not speaker-dependent.

We investigate the viability of using the DST as an efficient replacement for the Fast Fourier Transform in the field of Speech Recognition. We have already seen in Chapter 2 how effective the DST is in the isolation of time-frequency coefficients. By performing a summation over a time interval $[a,b]$ of the DST coefficients, thereby creating a histogram of the frequency content of the interval, we find that this histogram resembles the Fourier coefficients of that interval with remarkable accuracy. We shall call this histogram the *Spectrogram Dominant Scale Transform* (SDST). We tested this theory on several speech signals from the female speaker AKC, and the results are presented here.

For the accurate identification of any vowel sounds in speech, the identification of one or more of the formants used in the construction of the vowel is critical. In fact, three is often accepted as the minimum number of formants required. We investigate the possibility of using the DST and associated SDST as a formant detector of human vowel sounds.

Two techniques used successfully in the field of speech recognition are:

- Hidden Markov Models
- Cepstrum Analysis

3.1.1 Hidden Markov Models

Hidden Markov Models [Hol88] represent each word as a sequence of states, with transition probabilities between each state and its permitted successors, and probability distributions defining the expected observed features for each state. The model with the highest probability is assumed to represent the correct word.

3.1.2 Cepstrum Analysis

Cepstrum Analysis [Fla72, Sch75, Opp89] has been used to characterise speech. To compute the cepstrum, the Fourier Transform of the windowed speech time series is first computed. Windowing of the signal is accomplished by applying the Hamming window to the speech segment; this data window, 'D-window' in Figure 3-1, is given in Eq. 1-5. The logarithm is taken of the resulting magnitude spectrum. The inverse DFT of this signal is called the *cepstrum*. In speech applications, the cepstrum comprises two peaks for voiced signals; the pitch period being determined by the higher of the two. *Cepstral smoothing*, which is often used in speech analysis systems and for the measurement of formants, is accomplished by applying a low-time window, 'C-window' in Figure 3-1, to the spectrum. The smoothed spectrum retains peaks at the vocal tract resonances or formant frequencies.

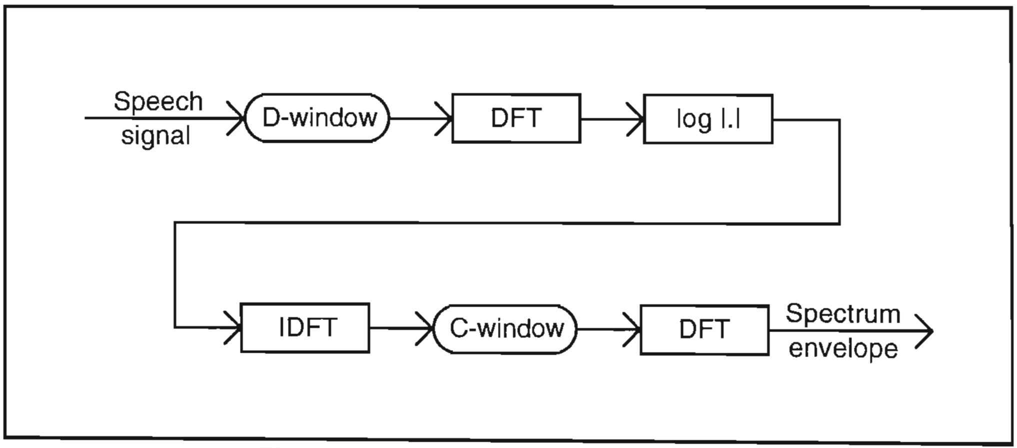


Figure 3-1: System for cepstrum analysis.

3.1.3 Dominant Scale Transform

The spatial and frequency localisation characteristics of the Dominant Scale Transform make it ideal for the decomposition of non-stationary signals. The human speech production mechanism consists of many co-ordinated parts and the resulting sound is dependent on a large number of variables. Speech sound waves are rapidly changing signals which are inherently complex. It was therefore somewhat of a challenge, using the DST, to venture into Speech Recognition. The author was greatly encouraged by the consistency of the results obtained using the DST with much of previous research from other authors in the field. The research was primarily concerned with obtaining a consistent model which allowed maximum speaker independence for word classification. Most of the work is concerned with the processing of the speech signal and extracting useful features. Speech recognition generally involves the following two processes: the selection and extraction of features of the selected speech signals, and decision making based on pre-defined class boundaries and the feature vector.

The initial conditions accomplishing this are very generalised in terms of speaker independence. We can divide the features into two distinct classes:

- *Speaker-independent features* such as silent passages, fricatives, and amplitude envelopes. While these characteristics do vary from person to person, there does remain a distinct consistency between speakers. Silent passages may vary in their time span, fricatives may differ where the emphasis is placed, and the amplitude envelope can, and probably will, differ from person to person; however, select generalities remain remarkably constant. A speaker-independent sub-language proposed by Dreyfus-graf [Dre72] was based on only three vowel categories and three consonant categories; namely /o/, /i/, /a/, and /s/, /t/, /n/ respectively. The SOTINA code for the digits "one" through "nine" is shown below. For example, 1993 would be pronounced "inanati". The SONITA coding system was found to be quite insensitive to foreign languages.

1	2	3	4	5	6	7	8	9
i	to	ti	ta	so	si	sa	ni	na

Table 3-1: The SONITA sub-language proposed by Dreyfus-graf.

- *Speaker-dependent features* are features such as formant frequencies, and other characteristics which are particular to a certain individual, or group of individuals. We try to preserve as many of the speaker-independent features as possible, since this would obviously allow for more reliable recognition across different races, sexes, nationalities, etc. We examine the speaker-dependent formant frequencies only after the word has been associated with a given class by the speaker-independent features.

3.2 Human Speech Production Mechanism

Although the purpose of this chapter certainly does not, and should not, entail a lesson of the Human Speech Production Mechanism, the author feels that the understanding of the processes involved in producing the complex speech signals, as well as their reception, is important in the construction of a speech recognition system. Moreover, knowing why a particular sound was made and under exactly what circumstances it will be repeated, lends insight into building a consistent recogniser. We therefore give a brief overview of the Human Speech Production Mechanism [Gre78].

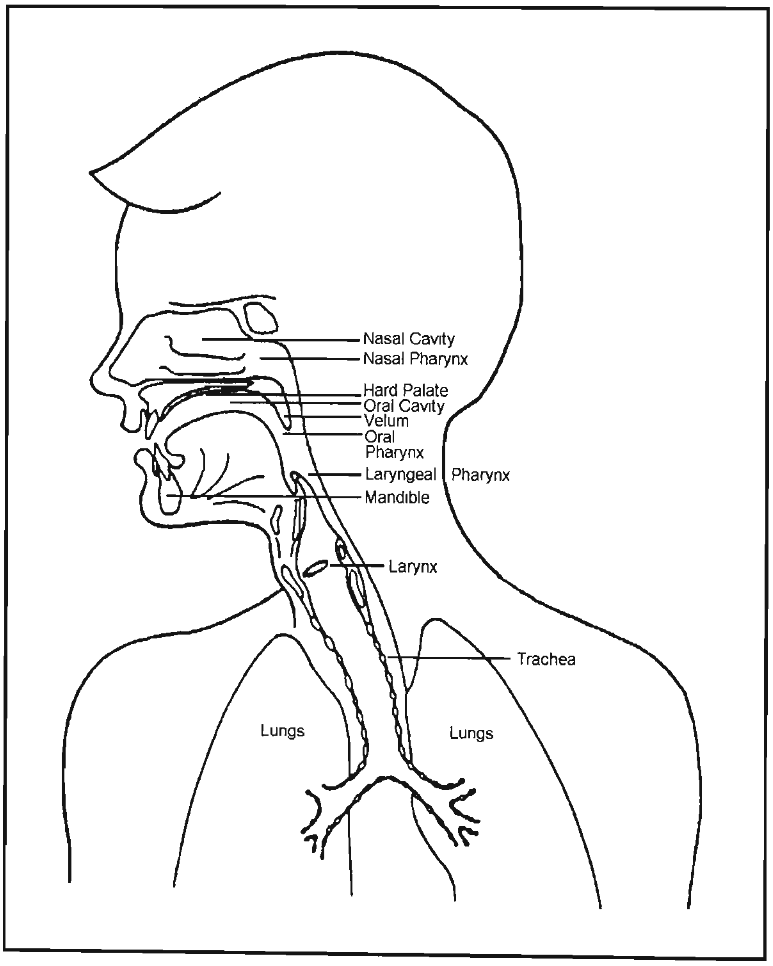


Figure 3-2: Human Speech Production Mechanism

Spoken language, or speech, is used as a natural means of communication between two people. Speech is transmitted through the air in the form of sound waves after being emitted primarily through the mouth and somewhat through the nose. The entire speech mechanism is a very well-co-ordinated instrument. The human apparatus concerned with speech production is complex and uses many important organs, namely, the lungs, mouth, nose, ears, and their controlling muscles. A good deal is known about the anatomy and physiology of speech production and perception, but less is known about the interaction of the brain with the vocal and auditory apparatus, although there are theories that attempt to explain the complexity of these interactions.

Speech sounds are produced when breath is exhaled from the lungs and causes either a vibration of the vocal chords (for vowels) or turbulence at some point of constriction in the vocal tract (for consonants). The sounds are affected by the shape of the vocal tract which influences the harmonics produced. The way in which the vocal cords are vibrated, the shape of the vocal tract or the site of constriction can all be varied in order to produce the wide range of speech sounds with which we are familiar.

3.2.1 Breathing

The use of exhaled breath is essential to the production of speech. In quiet breathing, of which we are not normally aware, inhalation is achieved by increasing the volume of the rib-cage. This reduces the air pressure in the lungs causing air from outside at higher pressure to enter the lungs. Expiration is achieved by relaxing the muscles used in inspiration so that the volume of the lungs is reduced due to the elasticity recoil of the muscles. The resulting increased air pressure in the lungs forces air out.

The form of expiration achieved by relaxing the inspiratory muscles cannot be controlled sufficiently to achieve speech or singing. For these activities, the inspiration muscles are used during exhalation to control lung pressure and prevent the lungs from collapsing suddenly. When the volume is reduced below that obtained by elastic recoil, expiratory muscles are used. Variations in speech intensity needed, for example, to stress certain words, are achieved by varying the pressure in the lungs. In this respect, speech differs from the production of a note sung at constant intensity.

3.2.2 The Larynx

There are two main methods by which speech sounds are produced. The first is called *voicing*, whereby the vocal cords, located in the larynx, are vibrated at a constant frequency by the air pressure from the lungs. The second gives rise to *unvoiced* sounds produced by turbulent flow of air at a constriction at one of a number of possible sites in the vocal tract. A schematic view of the larynx is shown in Figure 3-2.

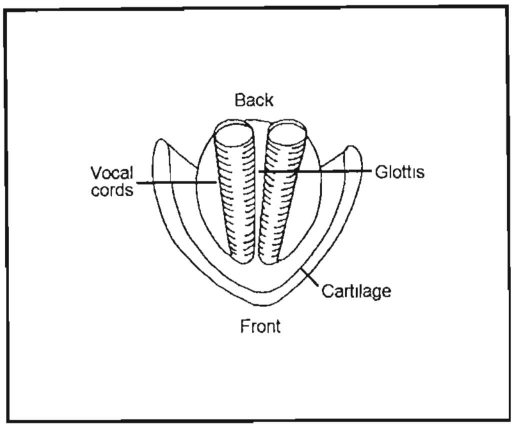


Figure 3-3: The larynx from above.

The vocal cords are at rest when open. Their tension and elasticity can be varied; they can be made thicker or thinner, shorter or longer and then can be either closed, open wide, or held in some position between. The accepted theory of *phonation*, or the production of voiced sounds, is called the *myoelastic aerodynamic theory of phonation* where "myo" refers to muscles. When the vocal cords are held together for voicing, they are pushed open for each glottal pulse by the air pressure from the lungs. Closing is due to the cords' natural elasticity and to a sudden drop in pressure between the cords, a result of the Bernoulli principle.

Just as the frequency of a plucked guitar string depends on the tautness and mass of the string, so too is the vibration frequency of the cords determined by the tension exerted by the muscles, and their mass and length. Men have cords between 17 mm and 24 mm in length, and those of women are between 13 mm and 17 mm. The average *fundamental* or voicing frequency, which is essentially the frequency of the glottal pulses, for men is about 125 Hz and for women about 200 Hz. For children it is usually more than 300 Hz. Figure 3-3 shows the range of fundamental frequencies produced by various singing voices and the frequency range of speech sounds. When the vocal cords vibrate, harmonics are produced at multiples of the fundamental frequency. The amplitude of the harmonics decreases with increasing frequency.

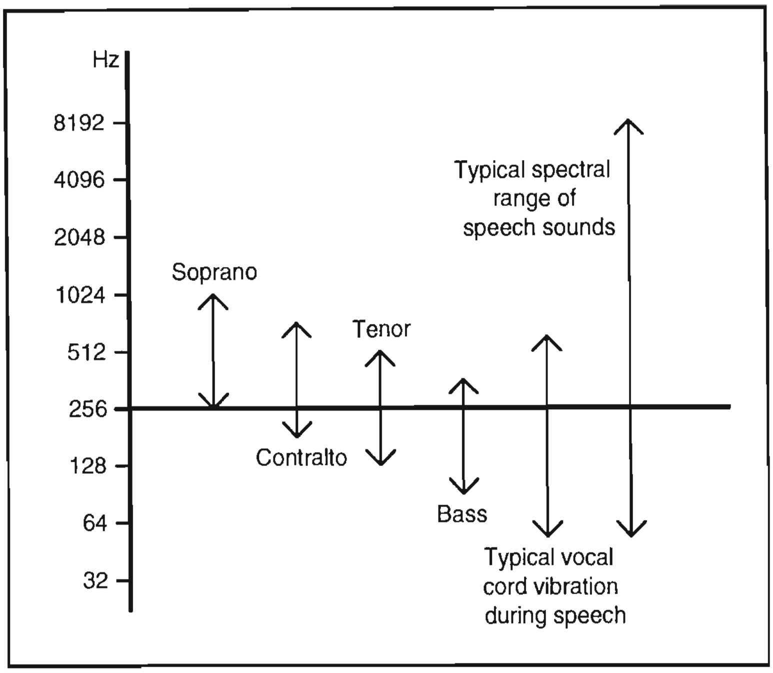


Figure 3-4: Frequency range of the human voice [Row92].

3.2.3 The vocal tract

For both voiced and unvoiced speech, sound that is emitted through the speaker's mouth and nose is a modification of the original vibration caused by the resonances of the vocal tract. The oral tract is highly mobile and the position of the tongue, pharynx, palate, lips, and jaw will all affect the speech sounds. The nasal tract is immobile, and is often coupled in to form part of the vocal tract depending on the position of the velum.

The tongue can move both up and down and forward and back, thus altering the shape of the vocal tract. It can also be used to constrict the tract for the production of consonants. By moving the lips outward, the length of the vocal tract can be increased. The nasal cavity is coupled in when the velum is opened for sounds such as /m/ in "hum". Here the vocal tract is closed at the lips and acts as a side branch resonator.

3.2.4 Acoustics of Speech Production

The vibration of the vocal cords in voicing produces sound at a sequence of frequencies, the natural harmonics, each of which is a multiple of the fundamental frequency. Our ears, however, judge the pitch of the sound from the fundamental frequency. The harmonics have reduced amplitude.

3.2.4.1 Formant frequencies

The tract responds to some of the basic and harmonic frequencies produced by the vocal cords better than others. For a particular position of the speech articulators, the lowest resonance is called the *first formant frequency* f_1, the next is the *second formant frequency* f_2 and so on. The formant frequencies for each of the vowel sounds are quite distinct, but each vowel sound generally has similar values regardless of who is speaking. For example, for a fundamental frequency of 100 Hz, harmonics will be produced at 200 Hz, 300 Hz, 400 Hz, 500 Hz etc. For the vowel /ee/ as in "he", typical values for f_1 and f_2 are 300 Hz and 2100 Hz respectively. The tongue is near the front of the mouth when making this sound and the high second formant results from the small size of the vocal tract cavity. For the vowel /ar/ as in "hard", typical values of the corresponding formant frequencies are about 700 and 900 Hz. The tongue is kept much flatter, and a much rounder sound is produced.

The fundamental frequency will vary depending on who is speaking, and on the speaker's mood and emphasis, but it is the magnitude and relationship of the formant frequencies which make each voiced sound easily recognisable.

3.2.5 Phoneme Production

The various categories of phonemes consist of vowels, diphthongs, semi-vowels, stop consonants, fricatives and affricates. We now examine a subset of these which we shall be dealing with later in the chapter.

3.2.5.1 Stops

Stop consonants or *plosives* are produced by forming a complete closure in the vocal tract, building up pressure from the lungs, and then suddenly releasing the pressure which is characterised by an explosion and aspiration of air. The point of constriction gives the specific speech sound. For /p/ the lips are held together, for /t/ the tongue is held against the alveolar ridge, and for /k/ the back of the tongue is raised towards the palate. The production of plosives may be modified by context; for example, the /p/ in "pot" emits much more air at the lips than in the modified version in "spot". Sounds of this kind occur at the beginning of the word "nine". If, in addition to the articulatory closure in the mouth, the soft palate is raised so that the nasal tract is blocked off, then the airstream will be completely obstructed, the pressure in the mouth built up, and an oral stop will be formed. When the articulators come apart, the air-stream will be released with a plosive quality. This kind of sound occurs in the word "two".

3.2.5.2 Vowels

We have already discussed the production of vowels in general as voiced speech sounds. Table 3-2 shows the formant frequencies of English vowels for typical male speakers.

Vowel	Example	Formant 1	Formant 2	Formant 3
/ee/	beat	280	2620	3380
/i/	bit	360	2220	2960
/e/	bet	600	2060	2840
/er/	bird	560	1480	2520
/ar/	father	740	1110	2640
/a/	hut	760	1370	2500
/u/	hood	480	740	2620
/uu/	loot	320	920	2200

Table 3-2: Formant frequencies of some English vowels for typical male speakers [Row92].

A simplified schematic picture for the /ee/ vowel is shown in Figure 3-4. The vowel quality comes mainly from the position of the tongue in the mouth. When producing the /ee/ sound, the tongue is moved forward and up to the roof of the mouth thus decreasing the size of the oral cavity. This produces high second and third formant frequencies which give the sound its characteristic tightness and can become squeaky when the speaker is stressed.

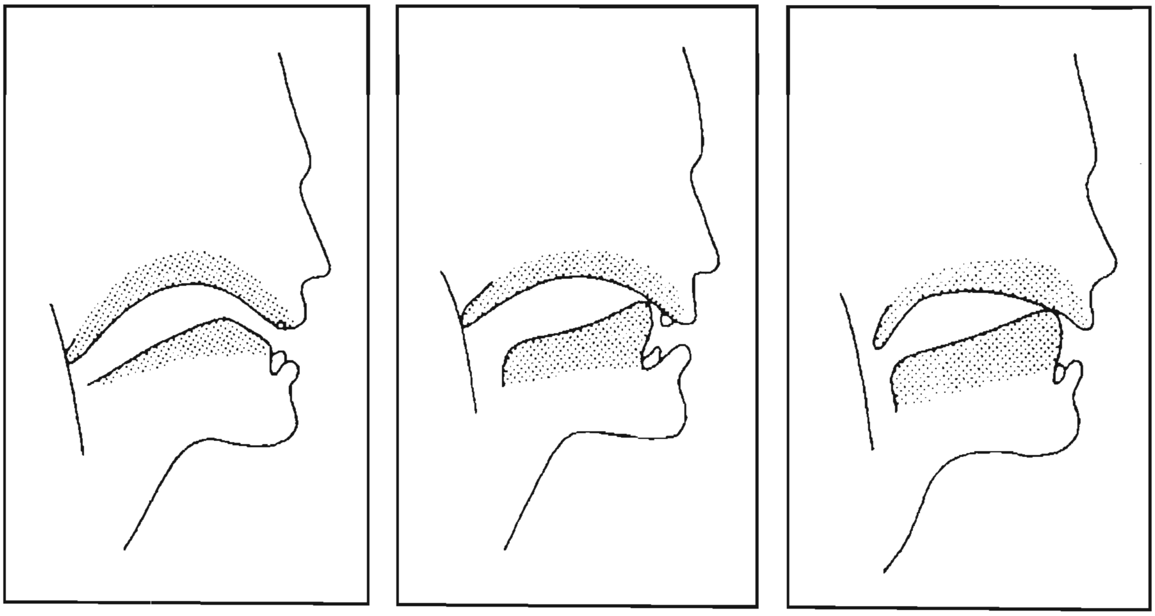
3.2.5.3 Nasals

The *nasals* /m/, /n/, and /ng/ are closely related to the stop consonants. However, there are some major differences. The distinctive sounds are produced by lowering the velum and making a closure in the vocal tract, thus introducing resonances in the nasal cavity with the oral cavity acting as a side branch resonator, while sound radiation is produced from the nostrils. Because the nasal passage is open, no pressure build-up occurs. The intensity of nasal sounds is lower than that of other speech sounds partly because the nasal cavity has very soft walls which absorb the sound.

3.2.5.4 Fricatives

Sustainable consonant sounds excited primarily by air turbulence are known as *fricatives*, and the turbulence referred to as *frication*. For example, regions of frication occur at the start of the word "three" and at the end of the "eight". In order to produce many of the familiar consonants, a constriction is formed at a point in the vocal tract and air is forced past creating friction and a very turbulent airflow, which causes a noisy random vibration. *Fricatives* may be unvoiced as for /f/, /th/, /s/ and /sh/, or can be combined with voicing which in combination with the same constrictions produces the four sounds /v/, /dh/, /z/, /xh/ respectively.

The constriction for /f/ and /v/ is formed by the lips, and for /th/ it is formed by the tongue pressing against the top teeth. For /s/, the tongue is pressed against the alveolar ridge and for /sh/ the tongue is held against the palate a bit further back than for /s/ in combination with a rounding of the lips. Thus the shape of the vocal tract for /f/ is similar to that for /p/, and the shape of the tract for /th/ is similar to that shown for /t/ in Figure 3-5.



Figures 3-5, 3-6, and 3-7: Typical vocal tract shape for /ee/, /t/, and /n/ vowel production.

3.3 The Physiology of the Ear

As we have seen, a speech signal is a complex combination of a variety of airborne pressure waveforms. This complex pattern must be detected by the human auditory system and decoded by the brain. The following section briefly looks at the construction of the human ear, and concludes by examining the close relationship between the human ear and the Wavelet Transform. We can consider the human ear to be divided into three parts:

- The *outer ear,* with its protective visible covering.
- The *middle ear,* which adjusts pressure levels between the outer and inner ear.
- The *inner ear* or *cochlea,* which contains the sensitive apparatus used to convert the sound energy into neural messages for the brain.

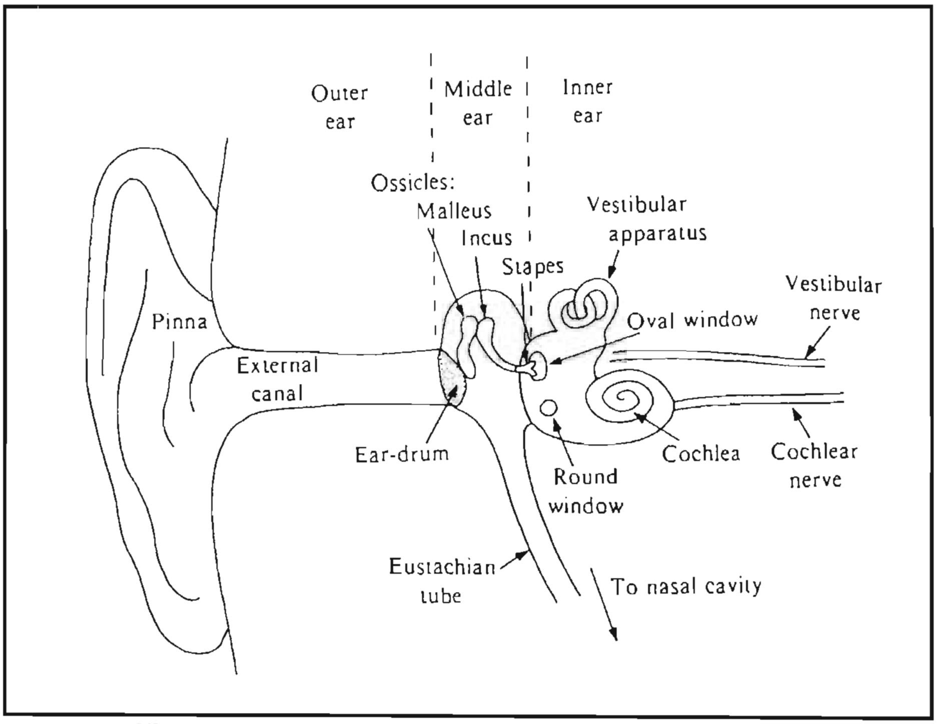


Figure 3-8: Schematic view of human ear (not to scale).

Sounds can be detected by both air conduction through the outer and middle ears and by bone conduction when vibrations travel through the bones of the head to the inner ear. Bone conduction is particularly important when a person is listening to his or her own voice.

3.3.1 The outer ear

The large visible outer part of the ear, called the *pinna*, offers protection. It also helps to focus the sound energy slightly, and because it is a little more receptive to sounds from the front of the head than from those behind, the pinna can help with the localisation of sound. It should, however, be noted that the primary means of localisation is by the timing and intensity differences between pressure waves arriving at the two ears.

The external auditory canal progresses from the pinna to the ear-drum (tympanic membrane) which separates the outer and middle ears. The canal is about 2.7 cm in length and causes a broad resonance effect which gives rise to an increase of sound pressure which is most effective between 2000 Hz and 5500 Hz, rising to a peak of about 12 dB at around 4000 Hz.

3.3.2 The middle ear

The middle ear overcomes air-to-liquid impedance to ensure that a detectable signal reaches the liquid-filled cochlea, to which it interfaces by two membranes called the oval window and the round window. Amplification is provided by a combination of two mechanisms. The first is by the action of the three small bones, or *ossicles*, which together act as a level to amplify pressure from the tympanic membrane to the oval window. These three bones, the malleus, incus, and stapes are named because of their shape (hammer, anvil, and stirrups respectively) and are the smallest bones in the human body! The stapes has a footplate of about 0.012 cm^2 which covers the oval window. The second amplification mechanism is provided by concentrating the pressure onto a smaller area, the tympanic membrane having an area about 18 times greater than the oval window. The overall effect is approximately a 30 dB pressure increase at the oval window over the pressure incident on the tympanic membrane.

3.3.3 The inner ear

The inner ear contains the cochlea which is of a coiled snail-like construction. When shown as though 'unrolled', the cochlea has a gradually tapering appearance from the base at the oval window to the apex.

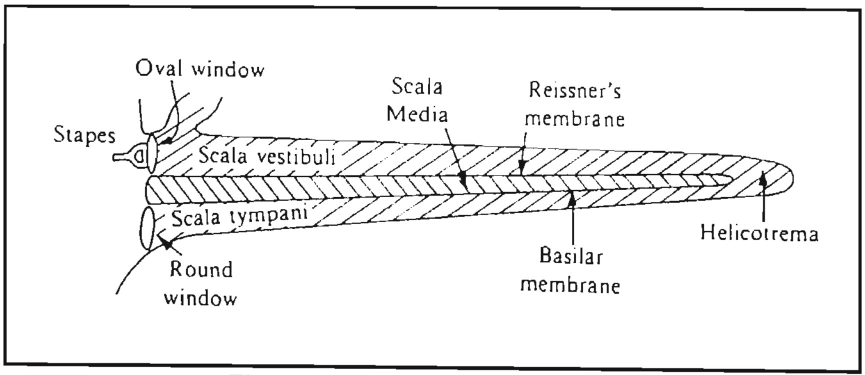


Figure 3-9: The 'unrolled' cochlea.

The cochlea consists of three principal fluid-containing enclosures which are separated by membranes. Pressure waves entering the cochlea at the oval window travel through the scala vestibuli along Reissner's membrane and through the narrow gap at the apex called the helicotrema to continue through the scala tympani along the basilar membrane to the round window where the pressure is released. The basilar membrane has a large number of tiny hair cells to which nerve endings are attached. The mechanical energy involved in the shearing of these hair cells by the pressure wave is transduced into the form of energy interpreted by the brain.

The pressure wave travels through the cochlea duct almost instantaneously so that the pressure difference occurs almost simultaneously at all places on the basilar membrane. The basilar membrane response is in the form of a travelling wave moving from base to apex whose amplitude increases gradually and then decreases rapidly. The peak of the pattern of vibration occurs at different places along the basilar membrane for different frequencies, being nearer the base where the membrane is narrow and stiff for higher frequencies and nearer the apex for lower frequencies. Thus the basilar membrane performs a kind of Fourier analysis on the incoming signal. For a pure tone, all parts of the membrane will vibrate with the same frequency, but some vibrate with a greater amplitude.

3.3.4 Does the ear perform a Wavelet Transform?

Our ear uses a wavelet transform when analysing sound, at least in the very first stage. The pressure amplitude oscillations are transmitted from the eardrum to the basilar membrane, which extends over the whole length of the cochlea. The cochlea is rolled up as a spiral inside our inner ear; imagine it unrolled to a straight segment, so that the basilar membrane is also stretched out. We can then introduce a co-ordinate y along this segment. Experimental and numerical simulation show that a pressure wave which is a pure tone, $f_\omega(t) = e^{i\omega t}$, leads to a response excitation along the basilar membrane which has the same frequency in time, but with an envelope in y, $F_\omega(t,y) = e^{i\omega t}\phi_\omega(y)$. In a first approximation, which turns out to be pretty good for frequencies ω above 500 Hz, the dependence on ω of $\phi_\omega(y)$ corresponds to a shift by log ω; there exists one function ϕ so that $\phi_\omega(y)$ is very close to $\phi(y - \log\omega)$. The proof is given in [Dau92].

The occurrence of the wavelet transform in the first stage of our own biological acoustical analysis suggests that wavelet-based methods for acoustical analysis promise to perform more suitably than other frequency analysis systems.

3.4 A CLOSE LOOK AT SPEECH WAVES

Now that we are aware of how spoken words are generated, we can proceed in earnest with the main purpose of this chapter, speech recognition. The chapter has the following structure:

1. Graphic plots of the digits (by speaker AKC) are presented and are discussed with reference to the human speech production mechanism. Distinctive characteristics of each of the signals are outlined.

2. We then investigate the Scale and Energy Functions of speech signals after the first pass of the Dominant Scale Algorithm. This provides very useful information in the segmentation of vowels and fricatives. Moreover, we show that the results from the Scale Function are totally consistent with research by other authors in the field.

3. Thirdly, we present what we feel is a significant result in terms of frequency analysis. For the vowel regions of speech, we show that the DST yields coefficients which match those of the Fourier Transform very closely . We use this fact to propose the use of the DST to generate spectra in vowel regions of human speech.

4. We conclude the chapter by presenting a speech recognition system which gives an excellent recognition rate for a very crude system which uses no 'intelligent' identification of the resulting spectra.

"One of the main difficulties of studying speech is that sounds are so fleeting and transient." [Lad73]

Our attempt at Speech Recognition of a limited vocabulary involved the English digits "one", "two", ..., "nine" using three speakers; two males and one female (DIC, HG, and AKC respectively). Each speaker was recorded saying each word ten times. The result was a database of 270 speech samples which were used throughout the research.

Time-domain signal of AKC's spoken word "One"

The word "one" begins with the /w/ semi-vowel in which the movement of the tongue is back, but more importantly, the lips are protruded and slightly closed. The latter attribute accounts for a slightly lower amplitude seen for $t = [0..0.04]$. The lips change position to generate the vowel /uh/ resulting in a sudden increase in amplitude which decays until the nasal /n/ is sounded by raising the tongue to the palate. Figure 3-10 illustrates a close up of the /uh/ vowel. The waveform is complex, but repeating.

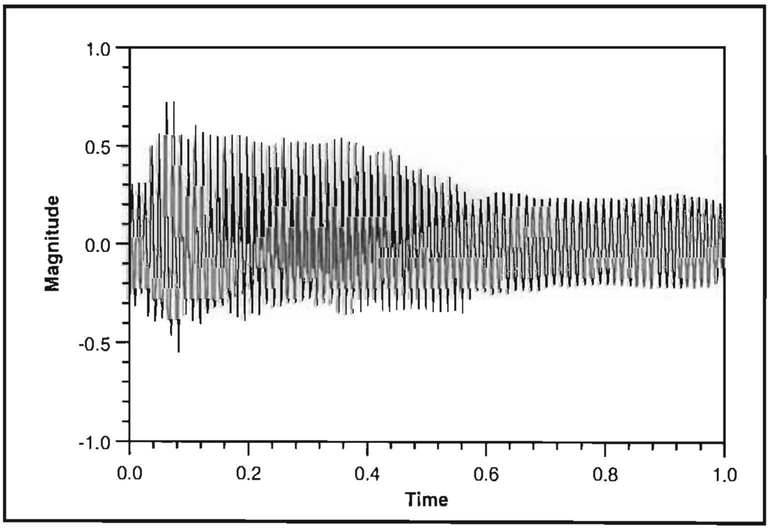


Figure 3-10: AKC saying "one".

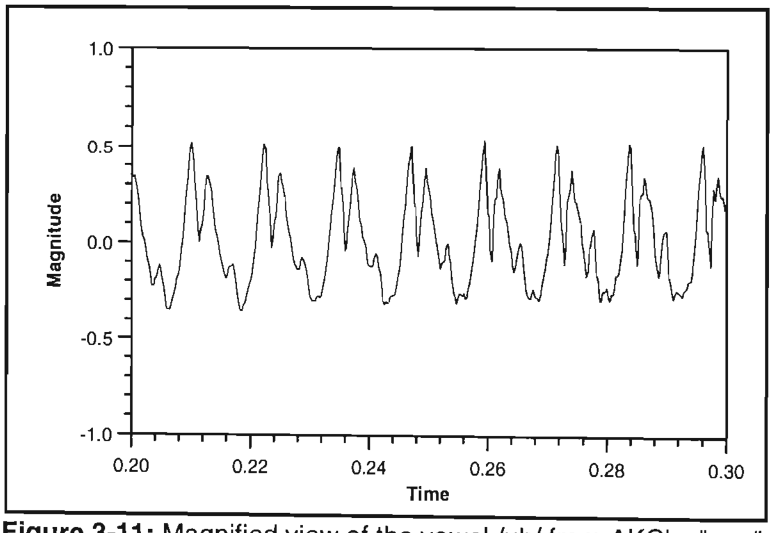


Figure 3-11: Magnified view of the vowel /uh/ from AKC's "one".

Time-domain signal of AKC's spoken word "Two"

The word "two" begins with the plosive /t/ which has the tongue pressed against the palate and then releasing an amount of air pronouncing the vowel /u/. For this vowel, the lengthening of the oral tract formed by the protruding of the lips and elevating of the tongue towards the back of the mouth lowers both the first and second formants. Figure 3-12 shows a close up of the /t/ fricative.

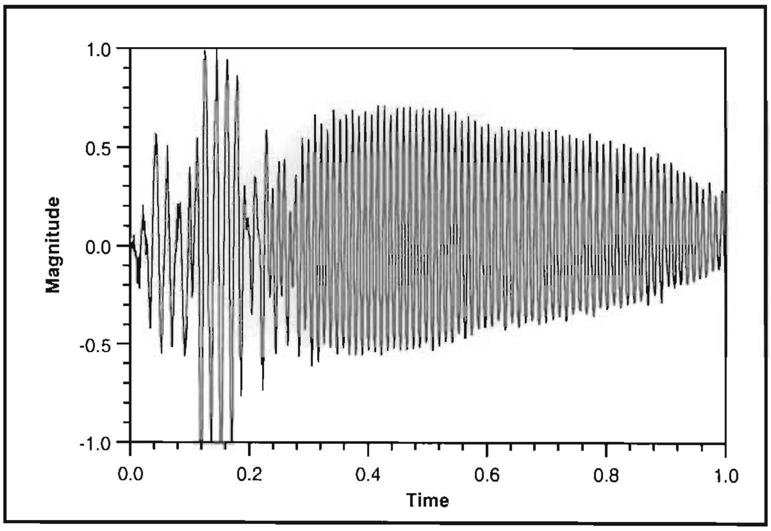


Figure 3-12: AKC saying "two".

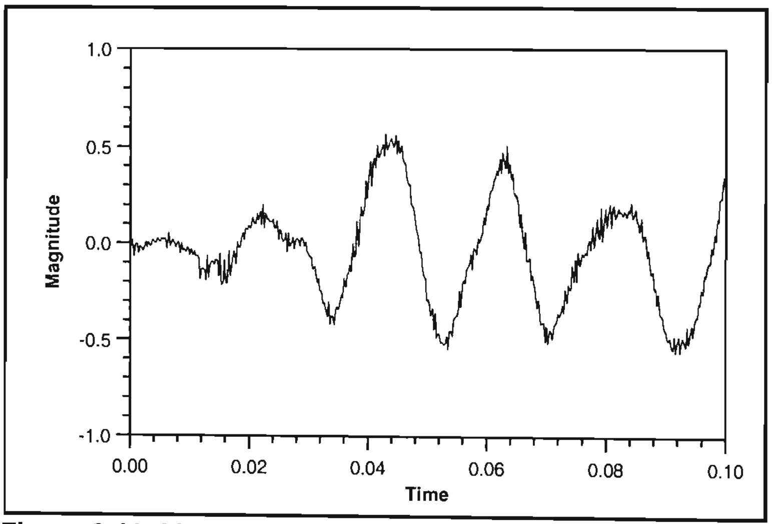


Figure 3-13: Magnified view of the fricative /t/ from AKC's "two".

Time-domain signal of AKC's spoken word "Three"

The word "three" begins with the fricative /th/ which has the tongue pressed against the teeth and causes low amplitude noise resulting from friction from the air. Note the brief pulse in amplitude around $t = [0.24..0.32]$, due to the sudden rush of air as the tongue is lowered from the palate. The semi-vowel /r/ follows and is produced principally from resonances in the vocal tract. The word ends with the decaying /ee/ vowel. This vowel is produced while the tongue is moved forward and up to the roof of the mouth, thereby decreasing the size of the oral cavity. Figure 3-14 shows a magnified view of the vowel /ee/.

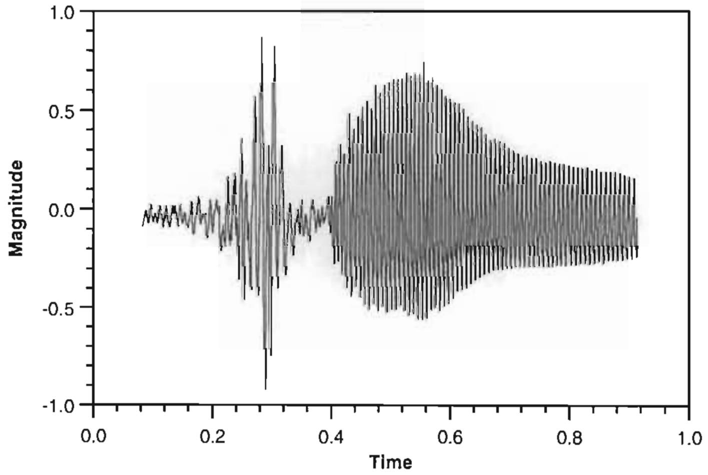


Figure 3-14: AKC saying "three".

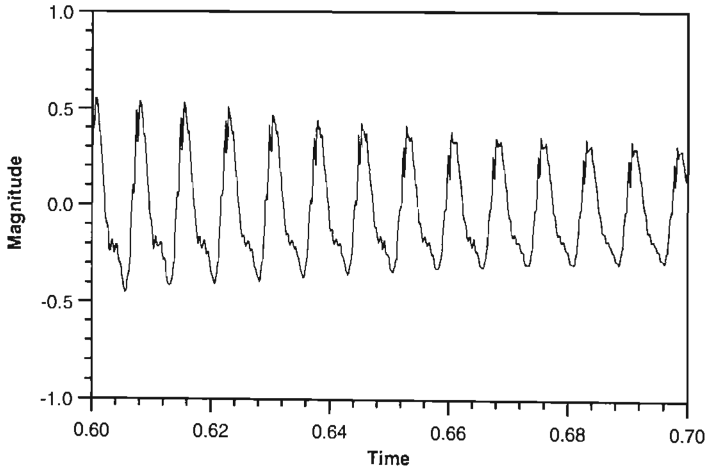


Figure 3-15: Magnified view of the vowel /ee/ is AKC's "three".

Time-domain signal of AKC's spoken word "Four"

The word "four" begins with the low amplitude unvoiced fricative /f/ created by a constriction caused by the lips. The vowel /aw/ succeeds the fricative and decays till the end of the word. The vowel /aw/ is magnified in Figure 3-16.

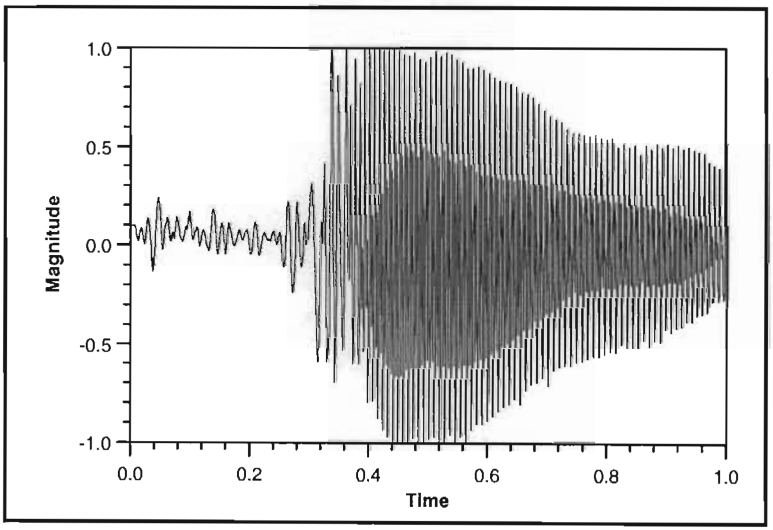


Figure 3-16: AKC saying "four".

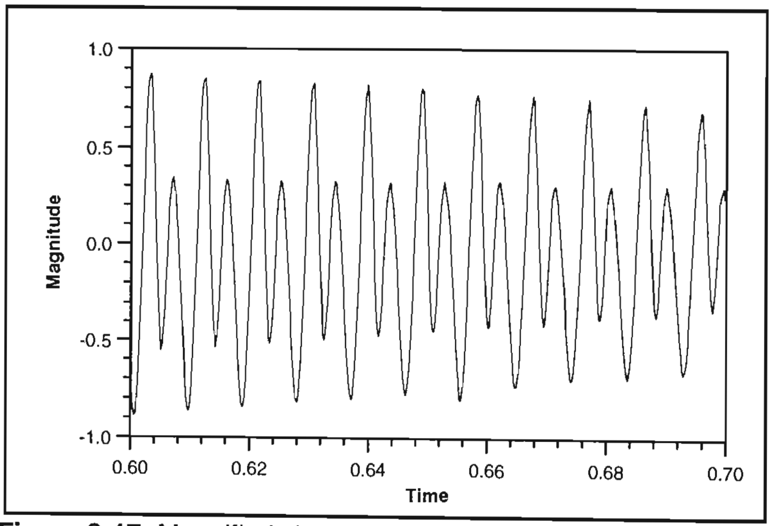


Figure 3-17: Magnified view of the vowel /aw/ from AKC's "four".

Time-domain signal of AKC's spoken word "Five"

The word "five" begins with the unvoiced fricative /f/ created by a constriction caused by the lips. The vowel /ie/ follows the fricative. The word concludes with the fricative /v/ caused by a constriction of the lips. Figure 3-18 shows a close up of the /ie/ vowel. Note the complexity of the waveform consisting of the fundamental as well as higher harmonics.

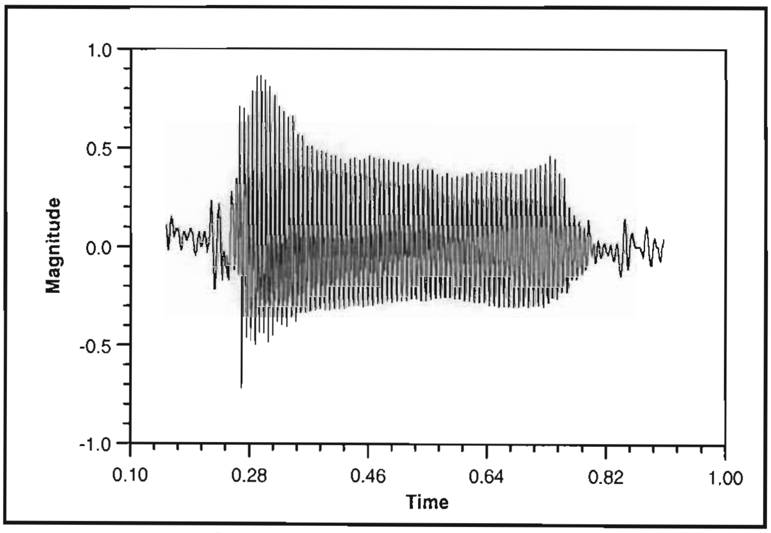


Figure 3-18: AKC saying "five".

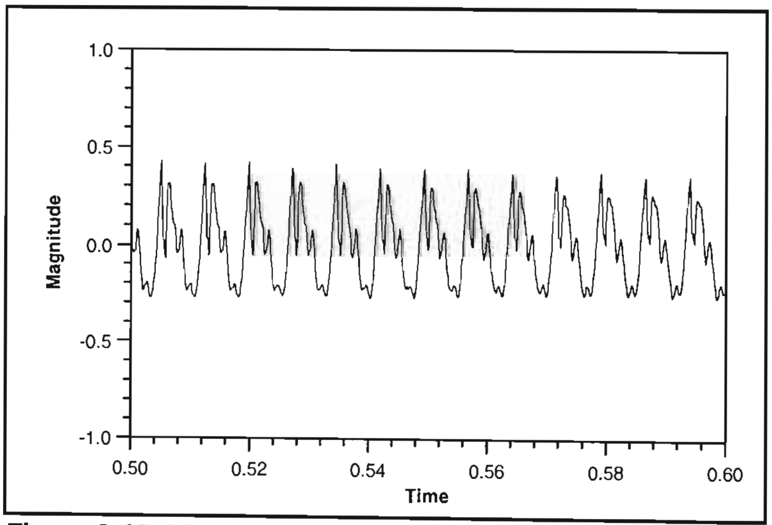


Figure 3-19: Magnified view of the vowel /ie/ from AKC's "five".

Time-domain signal of AKC's spoken word "Six"

The word "six" begins with the unvoiced fricative /s/ and the vowel /I/ follows that. A characteristic silent time lasts approximately 80 milliseconds between $t=[0.44..0.6]$ in Figure 3-19. The fricative /ks/ ends the word. Figure 3-20 shows a close up of the /I/ vowel. Note the complexity of the waveform consisting of the fundamental as well as higher harmonics. Also careful examination of the stop reveals the coarseness of an 8-bit sampling system. Using a 16-bit sampling data size would double the resolution, double the amount of data to be processed, but not increase the processing time as the same number of samples would still need to be processed.

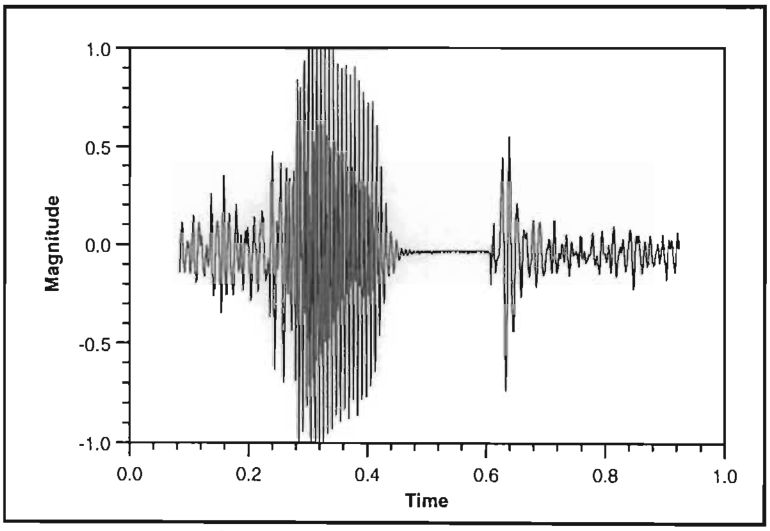


Figure 3-20: AKC saying "six".

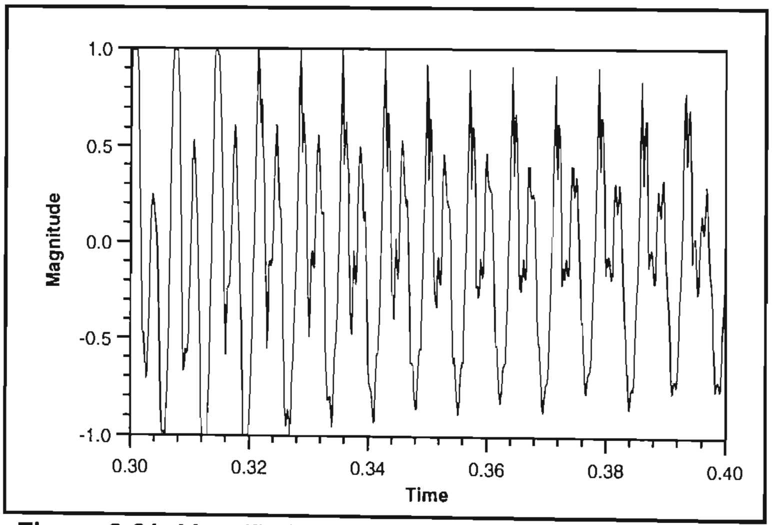


Figure 3-21: Magnified view of the vowel /I/ from AKC's "six".

Time-domain signal of AKC's spoken word "Seven"

The word "seven" begins with the unvoiced fricative /s/ and the vowel /ε/ follows. Then another fricative occurs; in this case, the voiced fricative /v/. Thereafter, the word ends with /I/ and the nasal /n/. Figure 3-22 shows a close up of the /I/ vowel. The word "seven" provides an excellent example of the different amplitudes with which the unvoiced fricatives, voiced fricatives, vowels, and nasals are spoken. Fricatives generally have lower amplitude and this is confirmed below with the unvoiced fricative /s/ on $t = [0..0.2]$, and the voiced fricative /v/ on $t = [0.44..0.56]$. The vowels are loudest appearing at $t = [0.24..0.44]$ and $t = [0.6..0.72]$.

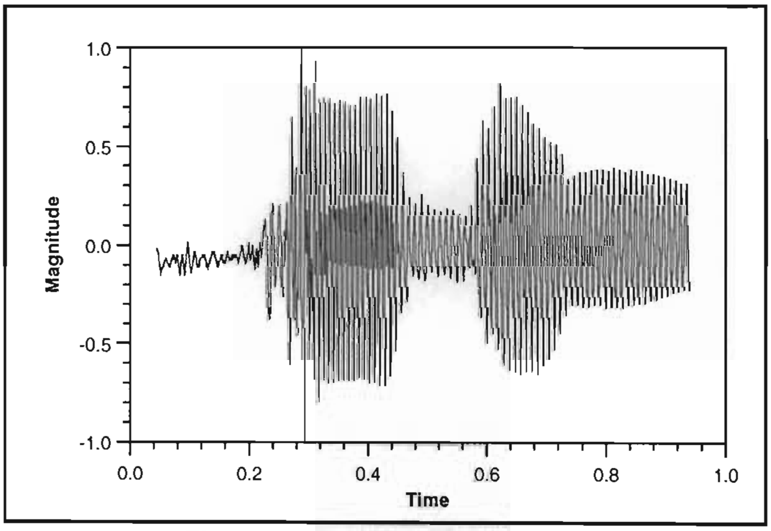


Figure 3-22: AKC saying "seven".

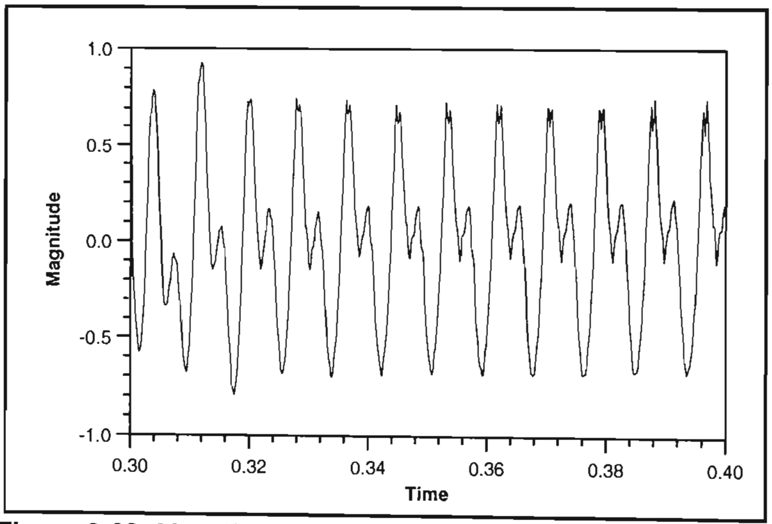


Figure 3-23: Magnified view of the vowel /ε/ from AKC's "seven".

Time-domain signal of AKC's spoken word "Eight"

The word "eight" begins with the vowel /eI/. This is terminated with a silent region of about 60 milliseconds. The word ends with the plosive /t/. The word "eight" is the only one in our data set having the format vowel-stop-fricative structure making detection relatively straight-forward. Figure 3-24 shows a close up of the /eI/ vowel. Again, note the complexity of the waveform and, in particular, the high frequency component.

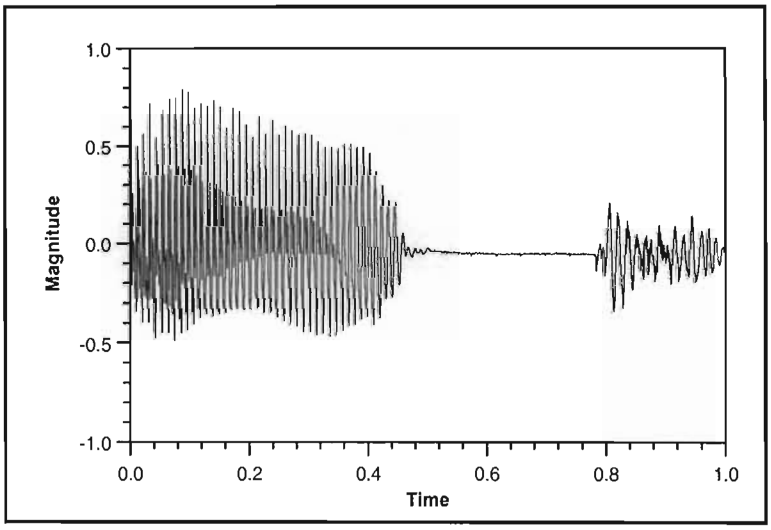


Figure 3-24: AKC saying "eight".

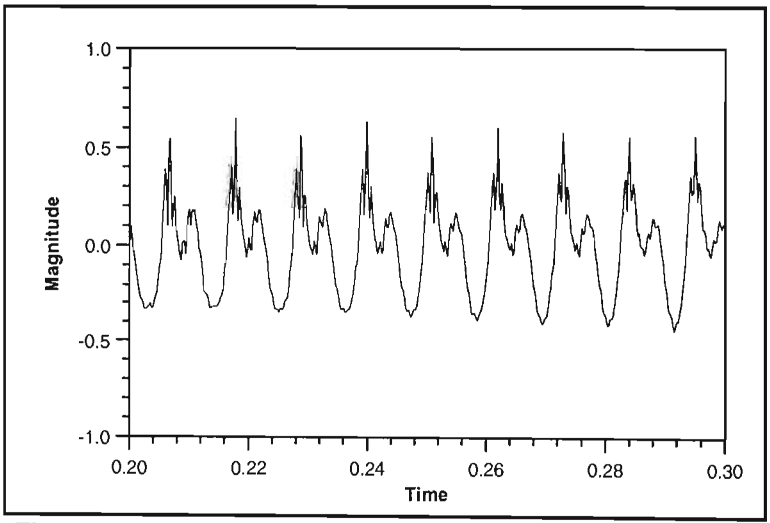


Figure 3-25: Magnified view of the vowel /eI/ is AKC's "eight".

Time-domain signal of AKC's spoken word "Nine"

The word "nine" begins and ends with the nasal /n/ created by lowering the velum and making a closure in the vocal tract, thus introducing resonances in the nasal cavity with the oral cavity acting as a side branch resonator. The sound is produced from the nostrils and the vowel /ie/ occurs between these two nasals. Figure 3-26 shows a magnified view of the /ie/ vowel. High frequencies are also present in this vowel.

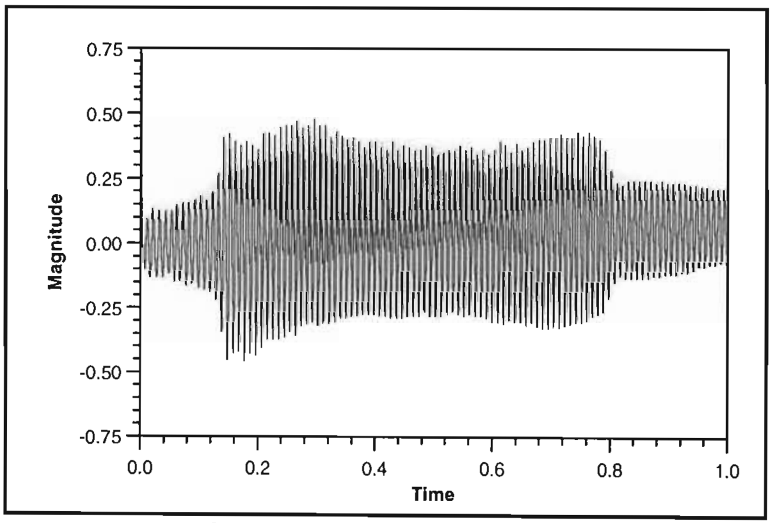


Figure 3-26: AKC saying "nine".

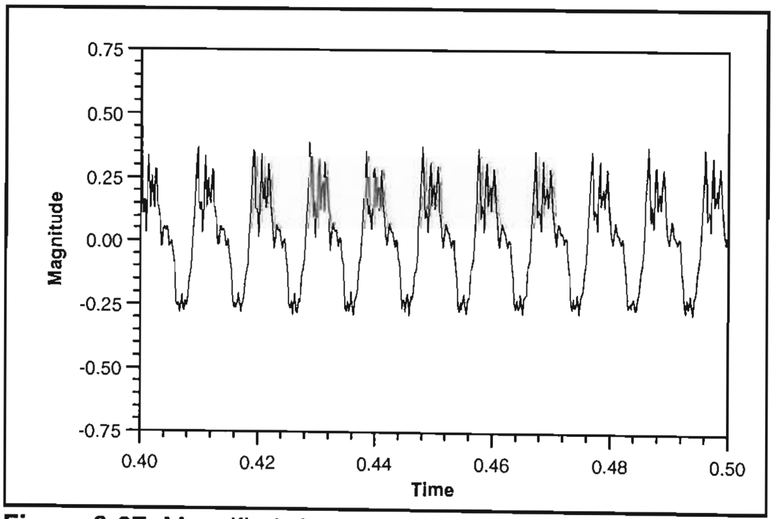


Figure 3-27: Magnified view of the vowel /ie/ from AKC's "nine".

3.5 SPEECH RECOGNITION USING THE DOMINANT SCALE TRANSFORM

The extraction of characteristics from each sample was to be used to provide us with a most probable favourite which we would report as the 'winner'. Speech signals represented in the time-domain appear extremely complex and reveal few readily identifiable characteristics. Vocal acoustic signals tend to have fairly characteristic properties which are better described in the frequency domain than the time domain. Practically all speech recognition systems, including this research, employ various strategies of extracting spectral information from speech signals. Spectral processing has several advantages. For example, consider a signal consisting of a fundamental and a higher harmonic. Now let us allow the phase of the harmonic to be set arbitrarily. The spectra would remain constant throughout even though the shape of the signal can change dramatically. What is interesting to note here is that a human's ear perceives a signal the same way irrespective of the phase of the components in a signal. Another reason for the processing of spectral information is the significant reduction of data to be processed. It goes without saying that the Windowed Fourier Transform is a powerful tool for feature abstraction. Its simple time domain to frequency domain mapping lets the system construct a time-frequency spectrogram. A newcomer on the scene is the Fast Wavelet Transform. It introduces an advantage in its octave-based time-frequency decomposition in that the same degree of spatial localisation is preserved. Unfortunately, as revealed in Chapter 2, we cannot trust the resulting coefficients with a significant level of confidence for spectral analysis.

3.5.1 The Fast Wavelet Transform and Speech Recognition

The Fast Wavelet Transform [Mal89b] is an obvious choice in the decomposing of speech signals for it generates both spatially and frequency localised coefficients which contain no redundancy due to the (assumed) use of orthogonal basis functions.

The author is of the opinion that, due to the coarse spatial localisation of the lower frequency basis functions, as well as the sparse frequency space, it would be unrealistic to expect consistent decomposition coefficients. This scepticism is in agreement with research carried out by a colleague [Mey93]. His results are summarised in Table 3-3.

Word	Accuracy
One	75%
Two	40%
Three	60%
Four	50%
Five	100%
Six	100%
Seven	50%
Eight	100%
Nine	50%

Table 3-3: Mey93 results using the FWT for speech recognition.

The recognition accuracy of the numbers "five", "six", and "eight" was 100%, but the accuracy of the numbers "one", "two", "three", "four", and "seven" averaged just above 50%. As anticipated, the Fast Wavelet Transform proved reliable when high frequencies were used in the identification of the number. For example, the number "eight" has lower frequencies at the start of the word, a silent region approximately half way through the pronunciation, and high frequencies at the end of the word. The determination process therefore primarily concerns itself with the silence, and occurrence or non-occurrence of high frequencies more than the interpreting of the entire time-frequency decomposition. It was therefore no surprise, and consistent with the author's hypothesis, that the Fast Wavelet Transform performed poorly overall. The failure of the FWT to perform successfully, encouraged us to evaluate the performance of the DST on speech samples.

3.5.2 The Dominant Scale Transform and Speech Recognition

The remainder of the chapter examines how both the Scale and Energy Functions calculated by the DST are used to achieve a high accuracy of isolated word recognition. We show the consistency with which the Scale Function allows accurate three-way segmentation [Gol93a] of the spoken word into: silence regions, regions of voiced speech, and regions of frication. These are the three most basic phonetic categories in human speech [Wai88]. This classification method allows us to make initial decisions of specific groupings of words.

We now begin to examine the characteristic features for which we shall be searching. These features are quite distinctive and have particular patterns occurring in the Scale and Energy Functions. The specific features are: stops, fricatives, and vowels.

3.5.2.1 The Detection of Stops

Stops occur in our word set in only two words; namely "six" and "eight" in the pronunciation of the /ks/ and /t/ fricative respectively. The fricative /ks/ is sustained whereas the /t/ is not sustained, being entirely a transient caused by a changing vocal tract condition. There is a short period of silence before the /t/, corresponding to the closed vocal tract, and then the transient, corresponding to a short burst of filtered noise as the vocal tract opens.

The identification of stops is a fairly trivial task which can be simplified even further. From Figure 3-27, it should be obvious that stop detection can be easily performed in the spatial domain. If the stop identification is to be performed in the spatial domain, it is important to be aware of two possible sources of problems:

- In Appendix A, we detail the sampling and storage strategies. For our research, 8-bit samples were used. Zero amplitude in this case was defined at 128 which is approximately the midpoint in the 8-bit range [0..255]. However, if a 16-bit resolution is used, then zero amplitude would be 0. Therefore the algorithm would be required to first determine which number system was in use to know exactly what represented zero amplitude.

- For the male speaker HG and DIC, the silence regions, or stops, were centred around 128 in our 8-bit signal. The signal exhibited very slight deviations from zero amplitude and always lay in the range [127..129]. The slight deviation from 128 could have occurred as a result of ambient noise, microphone noise, slight instability of the analogue-to-digital converter or many other reasons. It is nevertheless very constant and therefore easily detectable in the spatial domain. However, with the female speaker AKC, the magnitude of the signal during the stop deviated quite substantially from zero amplitude; see Figure 3-28. We are not quite sure what caused this phenomenon, but it occurred frequently throughout AKC's sampled words. Since $\int \psi(t) = 0$, the wavelet convolution considers differences within the wave, and any DC component is essentially ignored. It therefore makes sense to consider the energy function of the DSA rather than examining the signal in the spatial domain. The implicit ignoring of the DC offset is analogous to ignoring the zeroeth coefficient, or DC component, of a Windowed

Fourier Transform. Note that since the DST energy function would need to be calculated for future stages in the recognition process, no computation penalties are incurred.

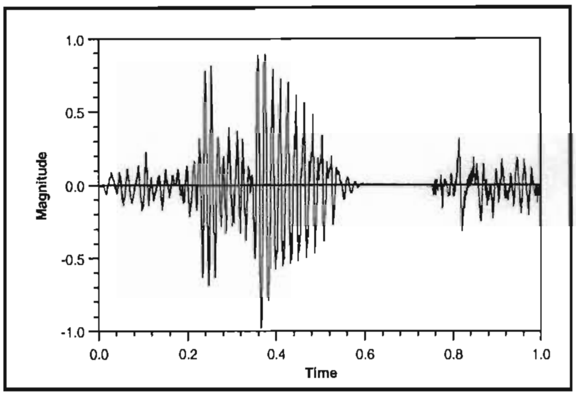


Figure 3-28: Speaker HG saying "six".

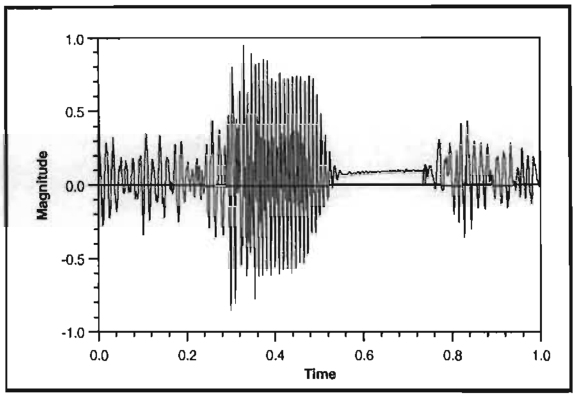


Figure 3-29: Speaker AKC saying "six".

Stop detection proved 100% accurate and this first stage of the recognition process allows us to divide the words into the following two classes:

Word	Contains a stop
1	No
2	No
3	No
4	No
5	No
6	Yes
7	No
8	Yes
9	No

Table 3-4: Using stops in the first class segmentation process.

3.5.2.2 The Detection of Fricatives

A major source of vocal acoustics is the air turbulence resulting from air being forced through a constriction in the vocal tract [Bad91, Scu92]; see Figure 3-5. These frication regions generally have lower energy than those regions containing vowels. When this constriction occurs between various parts of the tongue and the roof of the mouth, between the teeth, or between the lips, the air turbulence has a broad continuous spectrum [Hol88], or alternatively, it could be stated that the source has a flat spectrum [Ste72, Sol81]. It is this important fact which we shall exploit with the DST. When using cepstrum analysis, voiced speech is indicated by a distinct peak in the cepstrum at approximately 8 ms. Since unvoiced speech does not contain a periodic component, no strong peak appears in the cepstrum [Opp89].

We have already seen words ending in a stop followed by a region of frication; Figures 3-29 and 3-30 illustrate the words "three" and "eight", from the speaker AKC, which start with regions of frication.

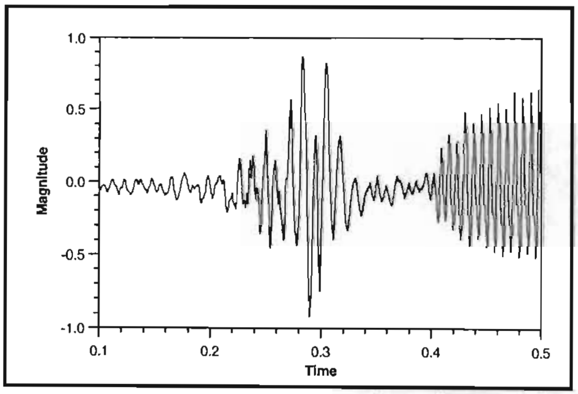


Figure 3-30: The /th/ at the start of the word "three" from the speaker AKC.

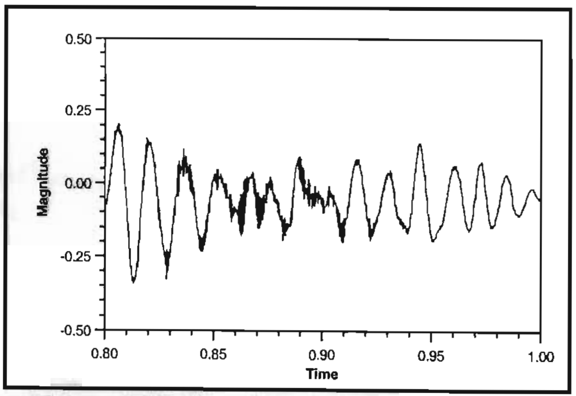


Figure 3-31: The /t/ at the end of the word "eight" from the speaker AKC.

The plosive sound of the fricative /th/ occurs in Figure 3-29 at around $t = 0.3$. This results from the constriction between the teeth and tongue being abruptly released. We found that this plosive occurred consistently with all three test speakers. In the following figures, examples of the Scale Functions for each digit are shown. The consistency with which the fricative regions induce an almost chaotic changing of the Scale Function corresponds exactly with the "broad continuous spectrum" findings of Holmes [Hol88]. Below each of the scale function plots, a measure of the volatility of the Scale Function is also shown. This function could be calculated in various ways. We examined two of the possibly many techniques. Each of the functions use a form of windowing. Although the function is calculated for each t_i for $i \in \{1..N\}$, in practice the calculations would be performed at large Δt intervals. Our first attempt used the standard deviation (STD) over the interval $[a..b]$ given by:

$$\sigma(t) = \left(\frac{\sum_{i \in \{a..b\}} (\alpha(t_i) - \alpha(\bar{t}))^2}{b-a+1} \right)^{1/2} \tag{3-1}$$

where $$\bar{t} = \frac{\sum_{i \in \{a..b\}} \alpha(t_i)}{b-a+1} \tag{3-2}$$

and α is the Scale Function. The results proved reliable, but computationally expensive. A more efficient technique was used which, for want of a better name, we shall refer to as the *Window Range* (WR). The WR is the difference between the largest scale and the smallest scale in the scaling function in the interval $[a..b]$ whilst ignoring the highest and lowest 5% to eliminate the possibility of one or two spurious values corrupting our calculation. The WR was often found to have slightly steeper gradients when entering or leaving vowel regions, or alternatively:

$$|\Delta RAF| > |\Delta STD| \tag{3-3}$$

For the sake of brevity, we abbreviate the window length for both the WR and STD functions as simply the *WR/STD Window Length.*

The broad frequency spectrum in the frication regions causes an almost chaotic changing of the DST coefficients, and consequently, a larger WR and STD. Additionally, some of these frication regions have high frequency components which reveal themselves as small scale wavelet coefficients. Interestingly, these small scale coefficients tend to appear consistently only in a subset of fricatives. The Scale and Energy Functions together with DST coefficients may therefore provide enough information to assist in the classification of fricatives. We leave this as an open problem. For our discussion on the choice of a suitable window length, we return to the ideas expressed in Chapter 2 and the problems associated with the Short-Time Fourier Transform. This time, however, the issues are slightly different:

Choosing the length of the WR/STD involves a calculated sacrifice of resolution, efficiency, and most importantly, accuracy. The decision can be summarised as follows:

- Longer WR/STD window lengths give rise to smoother WR/STD functions. This may be desired if absolute consistency is required, especially if Δt between WR/STD calculations is large. However, as the window length increases, so $|\Delta WR|$ and $|\Delta STD|$ decrease, in general, throughout the signal, which in turn tends to blur the transitions between the various regions.

- Shorter WR/STD window lengths tend to produce more erratic WR/STD functions, with the benefits of improved spatial localisation as well as improved efficiency.

The window length was found to be optimal (for our purposes) at approximately 60 ms. However, we must stress that this selection was based primarily on the choosing of the smallest window such that the WR/STD functions were relatively smooth. In a real-time application, we feel that this window length could be slightly reduced.

Fricative detection proved 100% accurate and this second stage of the recognition process allows us to further divide the words into classes.

Word	Contains a stop	Starts with fricative	Ends with fricative
1	No	No	No
2	No	Yes	No
3	No	Yes	No
4	No	Yes	No
5	No	Yes	Yes
6	Yes	Yes	Yes
7	No	Yes	No
8	Yes	No	Yes
9	No	No	No

Table 3-5: Using stops and fricatives in the second class segmentation process.

We now show the WR and STD plots for each of the digits we shall be examining. For all the following examples in the text, the speaker was AKC (except where otherwise noted). All the plots have been normalised relative to the largest value present in all the plots, thereby allowing a comparison of the magnitudes. We encourage the reader to compare these figures with earlier figures which show the time-domain signals of the digits.

3.5.2.3 Scale Limited Signals

Often, we wish to classify the range of the frequency components in a signal $f(t)$. This has importance in fields such as data sampling and data transmission.

Definition: A function $f(t)$ is *band-limited* with *bandwidth* 2Ω if its Fourier transform is equal to zero outside the interval $[-\Omega,\ \Omega]$; i.e.

$$F(\omega) = 0 \qquad \text{if} \qquad |\omega| \geq \Omega > 0$$

This well-known concept is based on the Fourier Transform. We modify it slightly to follow the course of our discussion regarding the DST.

Definition: The Scale Function α is said to be *scale-limited* over an interval $[a..b]$ if there exists a lower scale a_L and an upper scale a_U such that, for all $i \in \{a..b\}$:

$$\mathrm{a}_L \leq \alpha(i) \leq a_U$$

Clearly, any scale-limited signal is also band-limited. Throughout this chapter, we tend to use the term "scale-limited" rather loosely ignoring singularities exceeding either a_L or a_U. This is simply a matter of convenience.

Scale Function and WR/STD plots of "one"

Figure 3-31 shows the Scale Function of AKC's spoken word "one". The Scale Function is very well scale-limited throughout the signal. This accounts for the low WR/STD values seen in Figure 3-32. The lack of fricatives in the signal confine both the WR and STD to low values. Therefore, segmenting the words "one" and "nine" from the others involves searching for consistently low WR/STD values. The identification of vowel sounds, as we shall see later in the chapter, relies on the contributing frequencies, or formants, within the vowel regions. In Figure 3-32, notice that, in the nasal periods, between $t=\left[0..0.2\right]$ and $t=\left[0.6..1.0\right]$, the Scale Function does not venture into the smaller scale wavelets. However, for the vowel /uh/, the scale function does consist of these smaller support wavelets, thereby indicating a higher frequency content during that region. Note the sharp drop in (only) the WR function, at approximately $t=0.55$, where the vowel /uh/ is terminated.

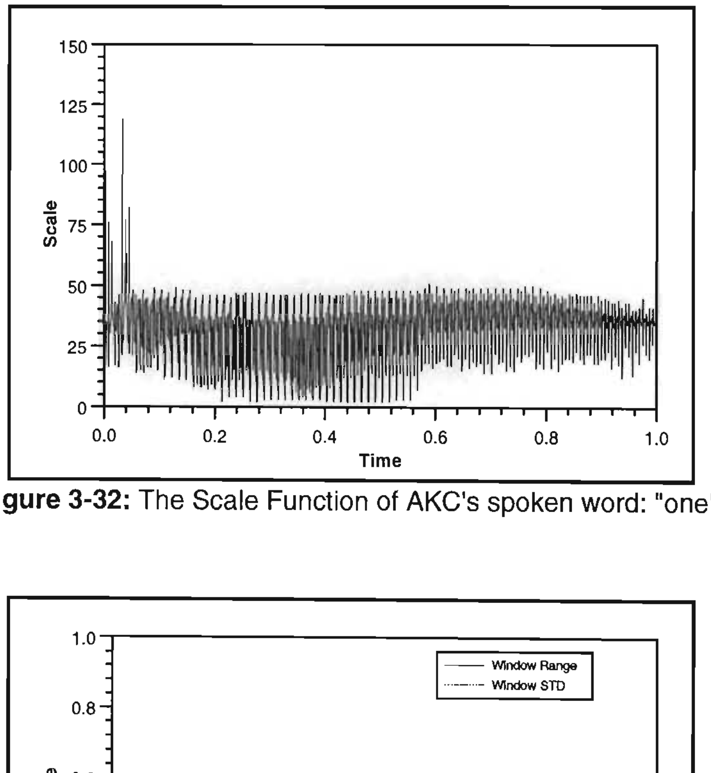


Figure 3-32: The Scale Function of AKC's spoken word: "one".

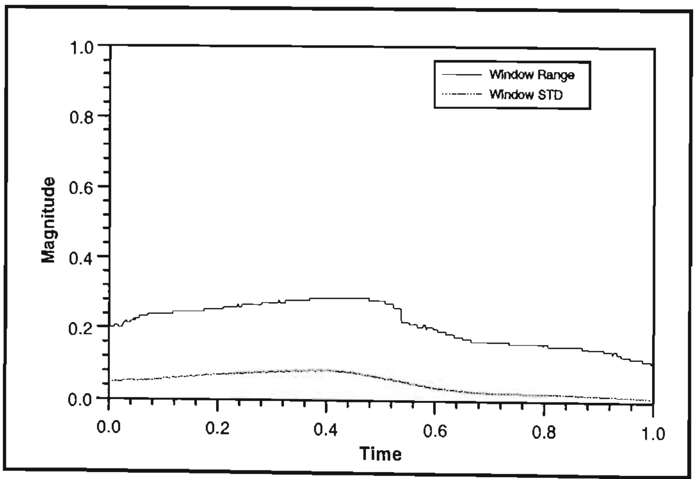


Figure 3-33: WR and STD plots for AKC's spoken word: "one".

Scale Function and WR/STD plots of "two"

Figure 3-33 shows the Scale Function of AKC's spoken word "two". The fricative /t/ at the start of the word causes the almost chaotic behaviour of the Scale Function for $t = [0.0..0.3]$. As we shall see for many of the other digits, this response is entirely consistent. After approximately $t = 0.3$, the vowel /u/ is encountered and the Scale Function is very much scale limited. Note, however, that there do exist several spikes, and this is the reason that the definition for the Window Range excludes the top 5% and bottom 5% of values. The rapid fluctuations during the period $t = [0.0..0.3]$ translates into higher WR and STD values; see Figure 3-34. The WR has a sharper gradient during the transition from the fricative to the vowel region. Obviously, the sharper this gradient, the more confident we can be of the transition.

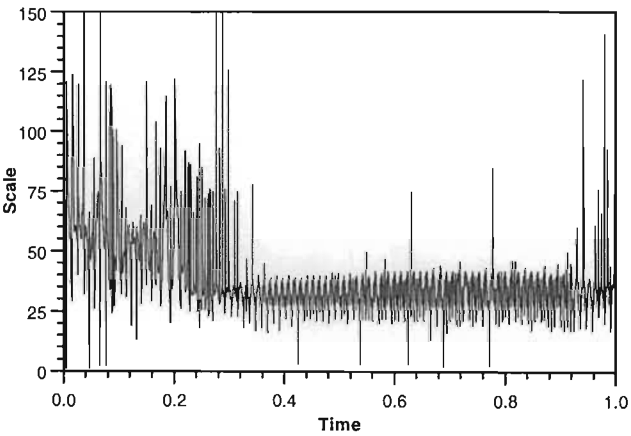


Figure 3-34: The Scale Function of AKC's spoken word: "two".

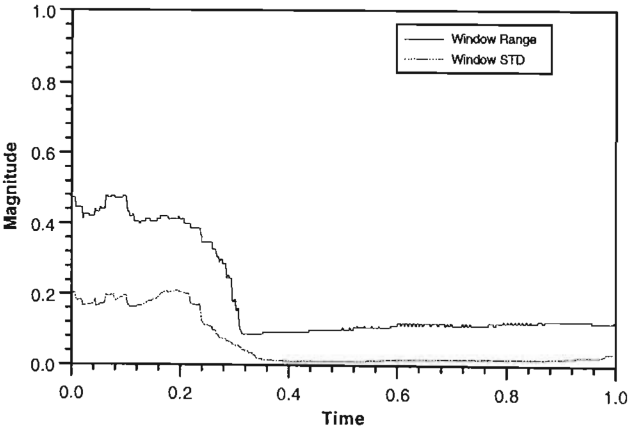


Figure 3-35: WR and STD plots for AKC's spoken word: "two".

Scale Function and WR/STD plots of "three"

Figure 3-35 shows the Scale Function of AKC's spoken word "three". The words "two", "three", and "four" are all quite distinctive in that they start with a fricative and the rest of the utterance is a vowel. A very encouraging characteristic of the Scale Function is its consistency regarding the chaotic nature of scales in the fricative regions, and then rapidly becoming scale-limited into the formant scales. We see clearly from Figure 3-35 that the fricative lies in $t = [0.0..0.4]$. The WR/STD values in the fricative regions also tend to be significantly higher than for the /t/ in "two". This is encouraging for future work regarding the identification and classification of fricatives.

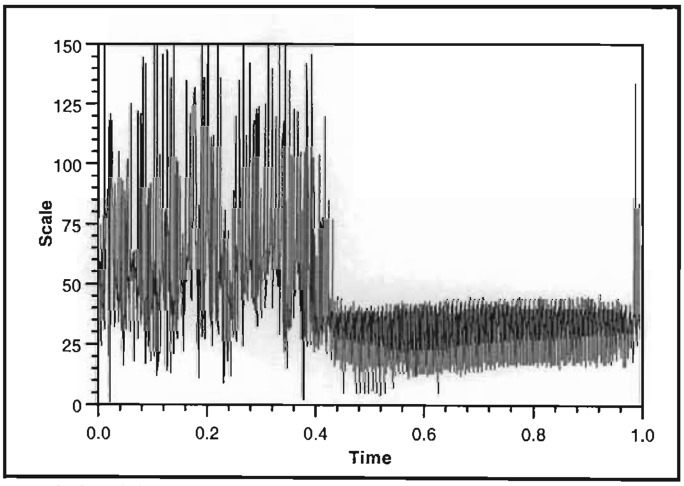


Figure 3-36: The Scale Function of AKC's spoken word: "three".

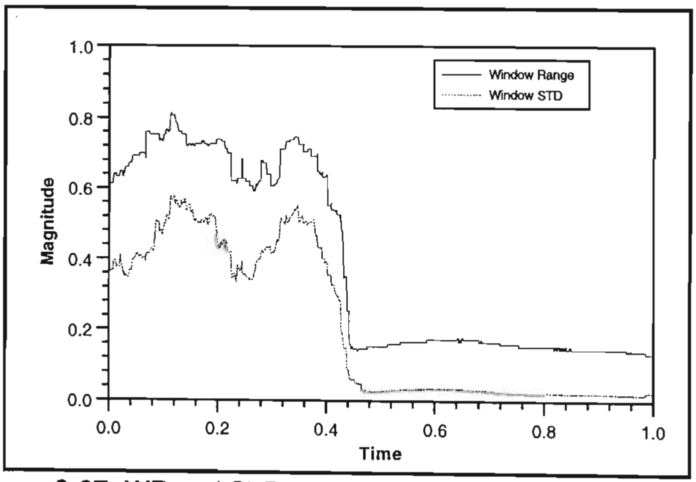


Figure 3-37: WR and STD plots for AKC's spoken word: "three".

Scale Function and WR/STD plots of "four"

Figure 3-37 shows the Scale Function of AKC's spoken word "four". Once again, the word begins with a fricative, this time /f/, and ends with a vowel. The fricative lies (approximately) in the region $t = [0.0..0.4]$. Research has shown that the fricative /f/ is also characterised by fairly high WR/STD values. As we have seen on the previous two pages, the basic fricative-vowel structure is common to the digits "two", "three", and "four". The classification can only be made once the vowel region has been examined more closely. This is dealt with later.

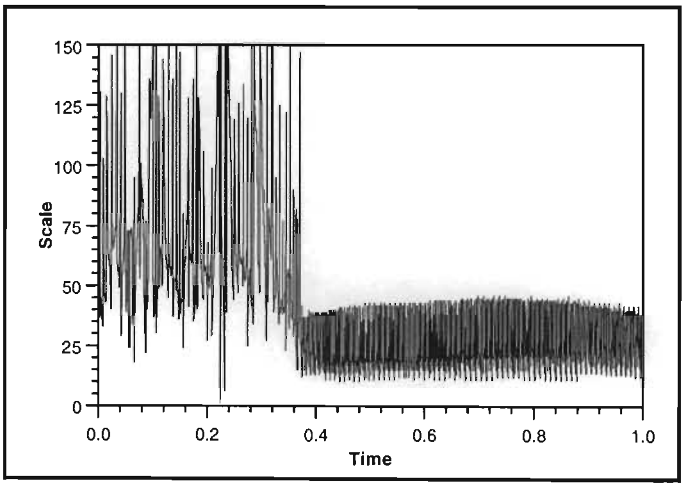


Figure 3-38: The Scale Function of AKC's spoken word: "four".

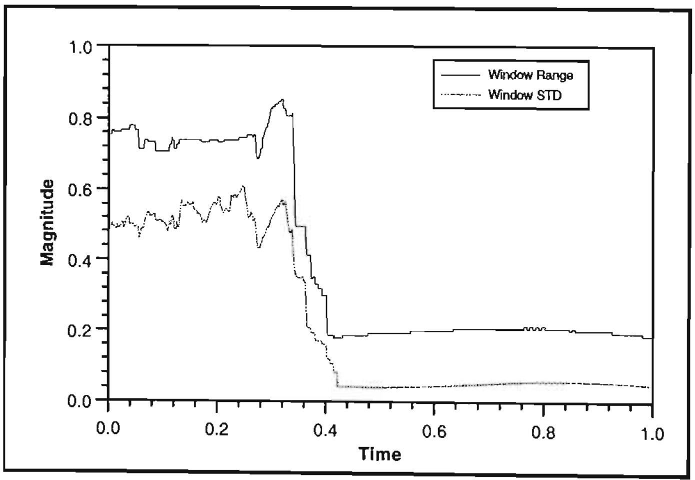


Figure 3-39: WR and STD plots for AKC's spoken word: "four".

Scale Function and WR/STD plots of "five"

Figure 3-39 shows the Scale Function of AKC's spoken word "five". The word "five" has the voiceless fricative /f/ and voiced fricative /v/ at the start and end of the word respectively. This causes the usual volatility of the Scale Function associated with fricatives. Figure 3-40 shows the U-shape we would expect. Therefore, our decision-making process searches for just two attributes; namely the U-shaped Scale Function, and no stops appearing during the utterance of the word. The word "six" has a very similar structure, but does include a stop. The two words are therefore distinguishable.

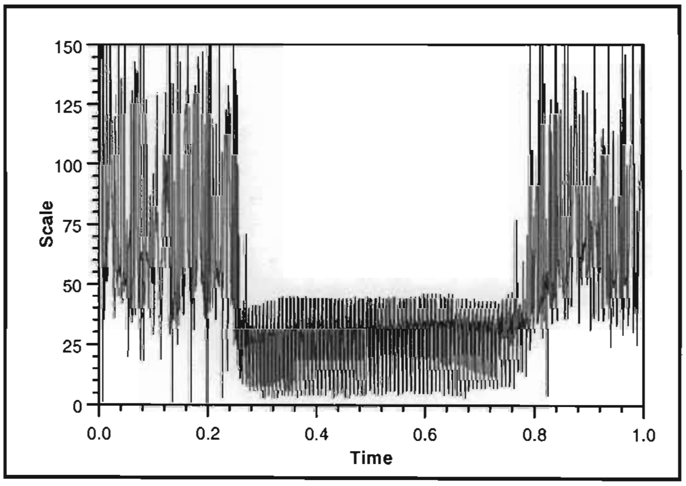


Figure 3-40: The Scale Function of AKC's spoken word: "five".

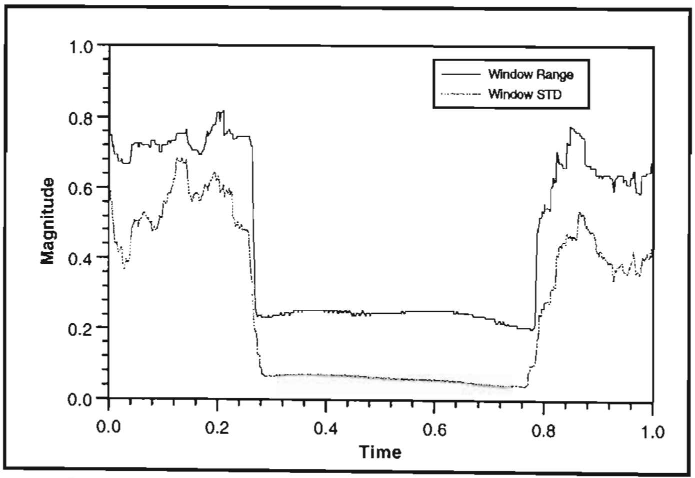


Figure 3-41: WR and STD plots for AKC's spoken word: "five".

Scale Function and WR/STD plots of "six"

Figure 3-41 shows the Scale Function of AKC's spoken word "six". The spoken word "six" exhibits the same U-shaped as the word "five". Although this may not seem obvious at first, it is worth noting the following characteristic about the DST: During stops, the Scale Function tends to fluctuate between the two extremes in the scale range. This is reflected in the WR/STD functions by very large values. Once a stop has been detected, we could simply ignore the WR/STD values in that region. Therefore, if we were to ignore the WR/STD values during the stop before the /ks/, the approximate U-shape is retained. Another fact worth noticing, is the very short time the vowel /I/ is uttered. In the example below, it lies on the small interval $t = [0.28..0.44]$. Low WR/STD values occur on this small time interval. The stop and U-shaped Scale Function are the two characteristics making the word identifiable.

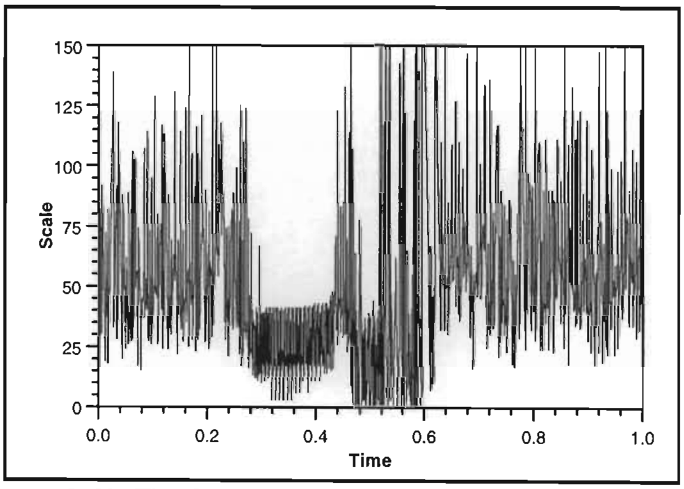


Figure 3-42: The Scale Function of AKC's spoken word: "six".

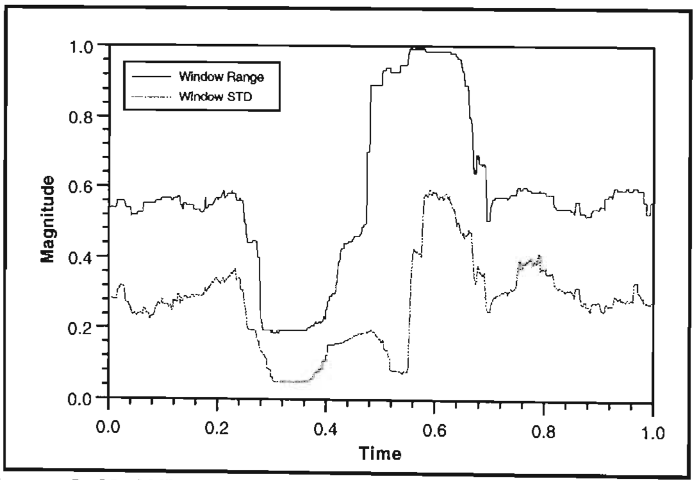


Figure 3-43: WR and STD plots for AKC's spoken word: "six".

Scale Function and WR/STD plots of "seven"

Figure 3-43 shows the Scale Function of AKC's spoken word "seven". The fricative at the start of the word "seven" is obvious and is clearly reflected in the WR/STD values. However, we made a very interesting observation with respect to all three speakers AKC, DIC, and HG. The voiced fricative /v/ occurring on the interval $t = [0.44..0.6]$ in Figure 3-44 does not exhibit the high WR/STD values it does at the end of the word "five". This finding is re-enforced later when we examine the DST of the word "seven". This result is totally consistent with [Lad85].

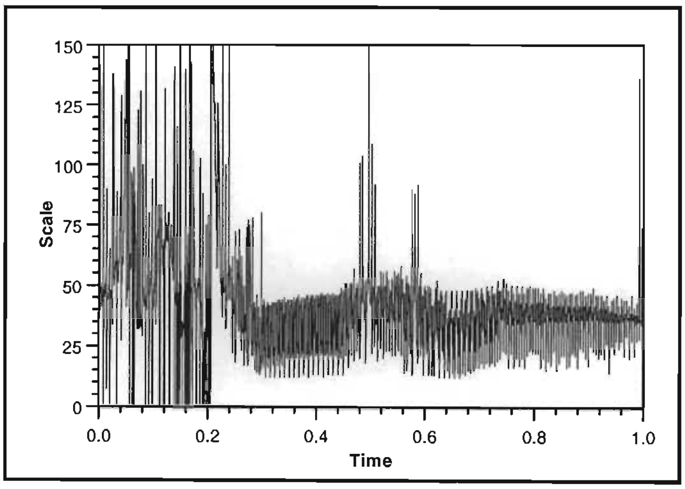


Figure 3-44: The Scale Function of AKC's spoken word: "seven".

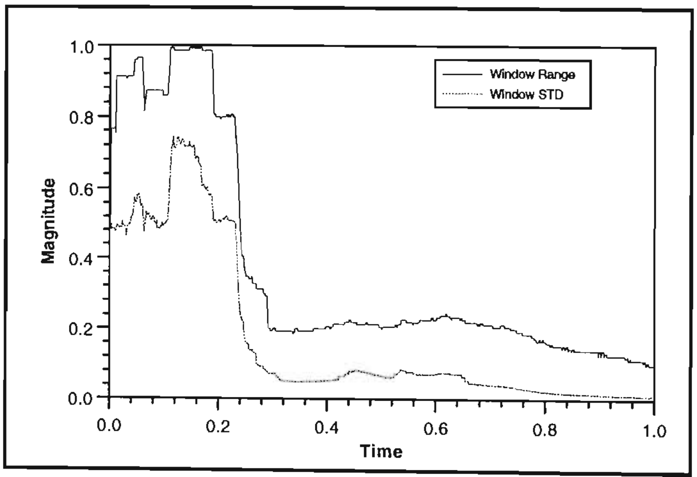


Figure 3-45: WR and STD plots for AKC's spoken word: "seven".

Scale Function and WR/STD plots of "eight"

Figure 3-45 shows the Scale Function of AKC's spoken word "eight". The spoken word "eight" exhibits characteristics similar to several of the other words, but is unique. Just like the word "six" in our data set, it has a stop, in this case on $t=[0.44..0.8]$. This stop, as in the case of "six", causes extremely high WR/STD values. As described earlier, these regions can be filtered out without much problem. The word starts with a vowel, which distinguishes it from the word "six". From close examination of Figure 3-23, we observe that the harmonics decay in magnitude and appear to have vanished for $t=[0.32..0.44]$. This event is reflected in Figure 3-45 by the lack of small scales in the Scale Function consistent with the lack of higher frequency harmonics.

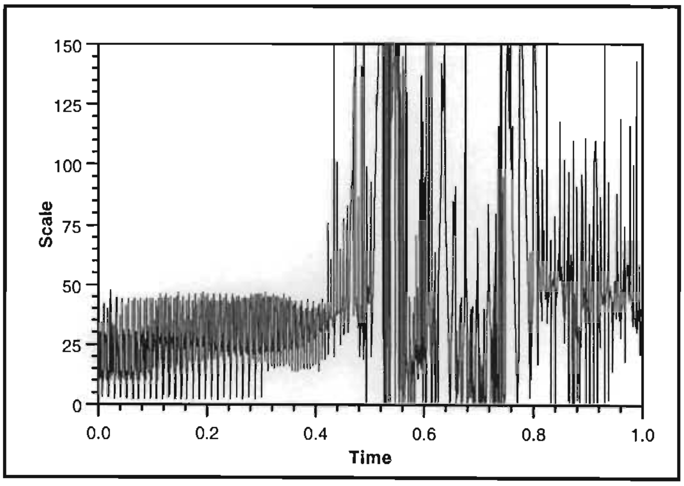


Figure 3-46: The Scale Function of AKC's spoken word: "eight".

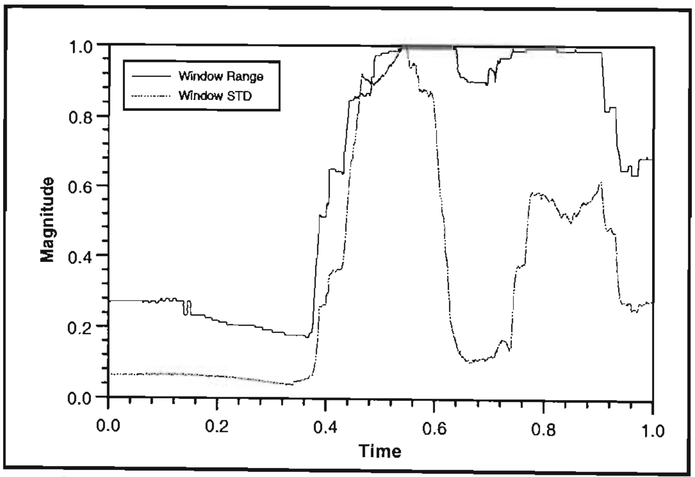


Figure 3-47: WR and STD plots for AKC's spoken word: "eight".

Scale Function and WR/STD plots of "nine"

Figure 3-47 shows the Scale Function of AKC's spoken word "nine". Similar to the word "one", the Scale Function is quite well scale-limited throughout the signal resulting in relatively low WR/STD values. The transitions between the nasal and vowel regions are clear from the WR plot. Although the fundamental appears relatively constant, some smaller scales are certainly present during the vowel sound on the interval $t = [0.16..0.7]$. The identification of vowel sounds, as we shall see later in the chapter, relies on the contributing frequencies, or formants, within the vowel regions. The nasal/vowel/nasal transition in the WR is even more pronounced in "nine" than in "one". This is barely noticeable in the STD plot. In contrast to the WR plot of the word "one", there exists a sharp rise of WR values upon entering the vowel from the nasal at approximately $t = 0.16$.

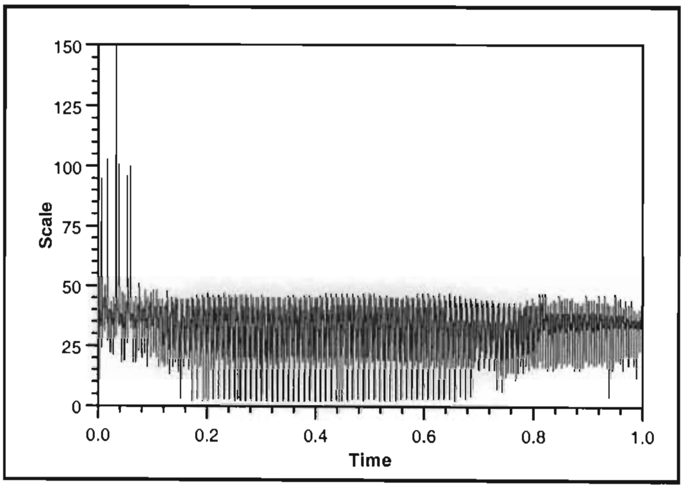


Figure 3-48: The Scale Function of AKC's spoken word: "nine".

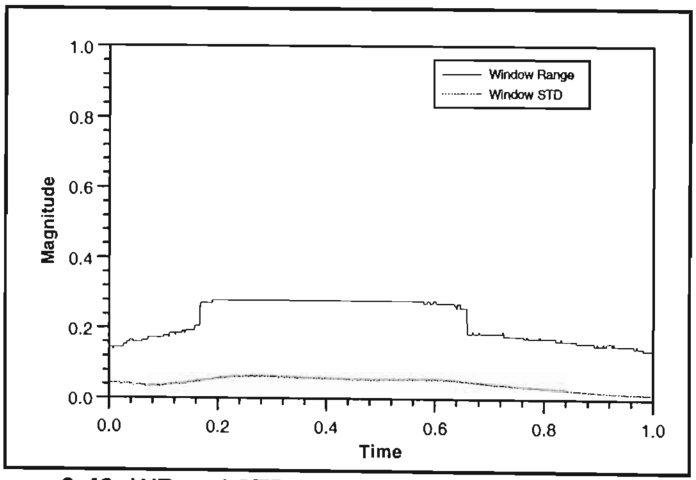


Figure 3-49: WR and STD plots for AKC's spoken word: "nine".

3.5.3 Time-Frequency Decompositions and Approximate Fourier Transforms of the Vowels

In this section, we present what we consider to be the most significant results of the text. The last few pages have been an examination of some of the properties of the Scale Function. We now turn our attention to the Time-Freq Function which is, in fact, a far more powerful function, and one that can be compared with established methods, such as the Fourier Transform. The Time-Freq Function decomposes a signal into its approximate time-frequency decomposition and we present the results of the words "one" to "nine". We shall show that in speech, the DST time-frequency space accurately delineates the changing formants throughout a spoken word. Speech compression using the time-frequency decomposition is currently being investigated by the author. Recall that the DST time-frequency coefficients are essentially localised at that component's zero-crossing point. Using this phenomenon, we can not only localise that particular component in frequency, but also in phase. This additional information can be utilised in many applications, some of which were mentioned earlier in the text.

3.5.3.1 Vowels and Formants

Nearly all vowel sounds are voiced, i.e., they are produced with vibrating vocal cords. Each time the vocal cords open and close, there is a pulse of air from the lungs. These pulses act like sharp taps on the air in the vocal tract, which is accordingly set into vibration in a way that is determined by its size and shape. In a vowel sound, the air in the vocal tract vibrates at three or four frequencies simultaneously. These frequencies are the resonant frequencies of that particular vocal tract shape. Irrespective of the fundamental frequency, which is determined by the rate of vibration of the vocal cords, the air in the vocal tract will resonate at three or four overtone frequencies as long as the position of the vocal organs remains the same. These resonances of the vocal tract are called *formants*. Individual differences are a result of different head sizes. A speaker with large vocal cavities will produce vowels with formant frequencies that are all lower than those of someone with a smaller head. Female speakers have formant frequencies that are, on average, 17% higher than those of males, some vowels being affected more than others because the formant frequencies are more affected by the comparatively large difference in pharynx size between males and females. In general, the formant frequencies of vowels spoken by speakers of the same dialect will retain the same relationship to one another, although they may by shifted up or down the frequency scale quite considerably relative to other speakers. Because of this, the formant frequencies at a single moment in time are not a good indication of vowel quality.

We have already shown how, using the Energy Function of the DST, we can segment signals into the three primary regions. We now focus on using the DST to identify the formants of the vowel region after the region has been located using the Scale Function. We have discovered what we feel is a fairly significant fact. By performing a summation of the DST coefficients over the interval $[a,b]$ and thereby creating a histogram of the frequency content of that interval, we find that this histogram closely resembles the Fourier coefficients of that interval. We shall call the histogram of the DST coefficients the *Spectrogram Dominant Scale Transform* (SDST). The following time-frequency decompositions are accompanied by SDST plots superimposed on the Fourier coefficients of that interval. The coefficients were scaled and normalised appropriately and relative values are therefore comparable.

We now examine the Time-Freq Functions and SDSTs generated from the words "one" to "nine" from the speaker AKC. The reader is strongly encouraged to continually refer back to the original time-domain plots.

DST and SDST function plots of "one"

Figure 3-49 shows the DST of AKC's spoken word "one". Careful examination of the DST shows that it is possible to segment the word "one" into the vowels /w/ and /uh/ and the nasal /n/. As the lips are retracted from the /w/ vowel to the /uh/ vowel, the coefficients in the transition have diminishing scale correlating to the increasing frequency. The vowel /uh/ is centred around $t = 0.4$ with the nasal starting at approximately $t = 0.6$.

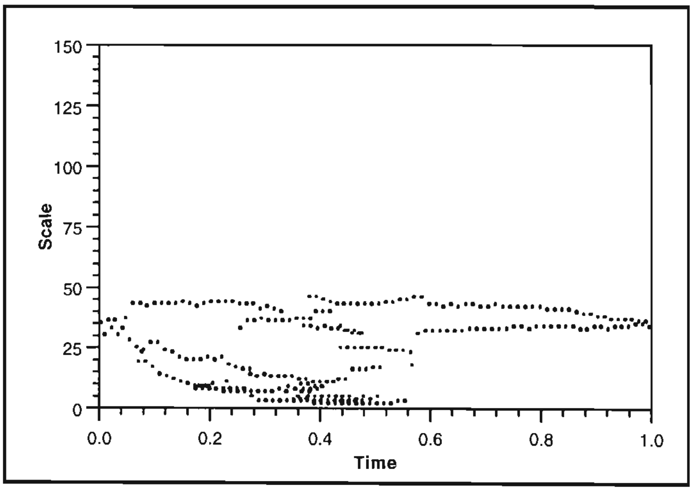


Figure 3-50: The time-frequency decomposition (DST) of AKC's spoken word: "one".

Figure 3-50 shows the Fourier coefficients superimposed on the SDST coefficients for the vowel /uh/. Most importantly, the fundamental is accurately represented. As a result of the rectangular windowing employed to perform the Fourier Transform, side lobes off the main lobe containing the fundamental frequency are clearly visible in the Fourier coefficients.

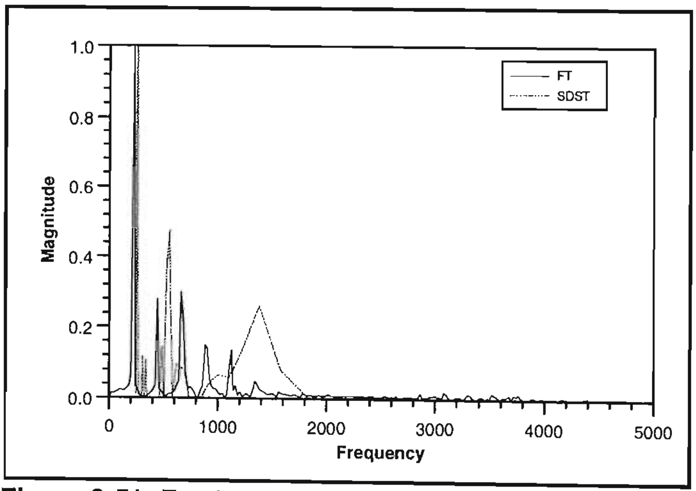


Figure 3-51: Fourier coefficients versus SDST coefficients for the vowel /uh/ of AKC's spoken word: "one".

DST and SDST function plots of "two"

Figure 3-51 shows the DST of AKC's spoken word "two". The fricative /t/ occurs during $t = [0..0.28]$ and has the associated rapidly changing DST coefficients. The vowel /u/ consists predominantly of a single sinusoidal and this is reflected in Figure 3-51 by a region of very constant DST coefficients ranging from $t = [0.28..1]$.

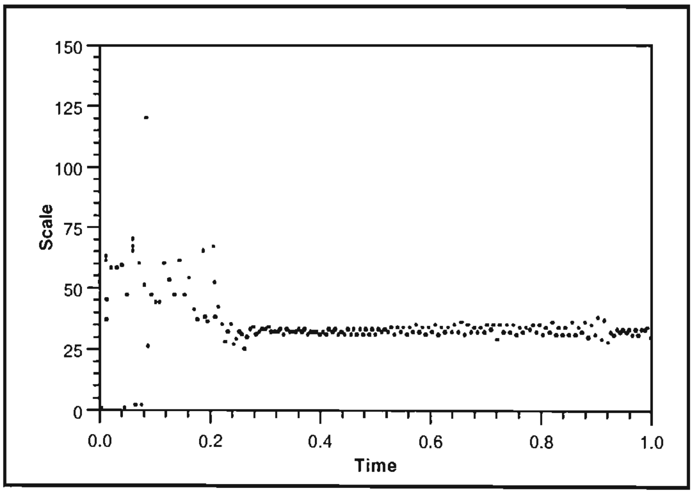


Figure 3-52: The time-frequency decomposition (DST) of AKC's spoken word: "two".

Both the Fourier and SDST coefficients represent a signal with a predominant sinusoidal. The peaks of the two transforms, in this case, are not equal and differ by approximately 80 Hz. The undersampling of the DSA accounts for approximately 30 Hz; therefore, we have a small discrepancy of only 50 Hz.

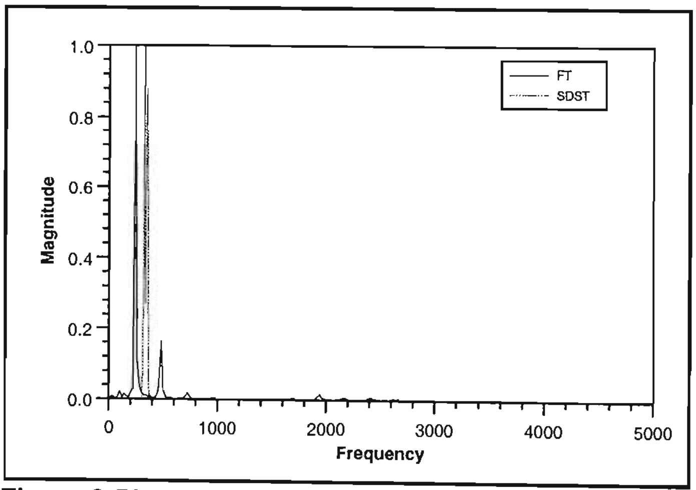


Figure 3-53: Fourier coefficients versus SDST coefficients for the vowel /u/ of AKC's spoken word: "two".

DST and SDST function plots of "three"

Figure 3-53 shows the DST of AKC's spoken word "three". The fricative /th/ occurs during $t = [0..0.4]$ and has the associated rapidly changing DST coefficients. The vowel /ee/ can be seen to have two distinct components. Interestingly, there exist higher frequency harmonics between $t = [0.52..0.64]$. The author is unsure exactly what in the Human Speech Production Mechanism produces these smaller scale coefficients. The FT probably would not have detected these short-time perturbations.

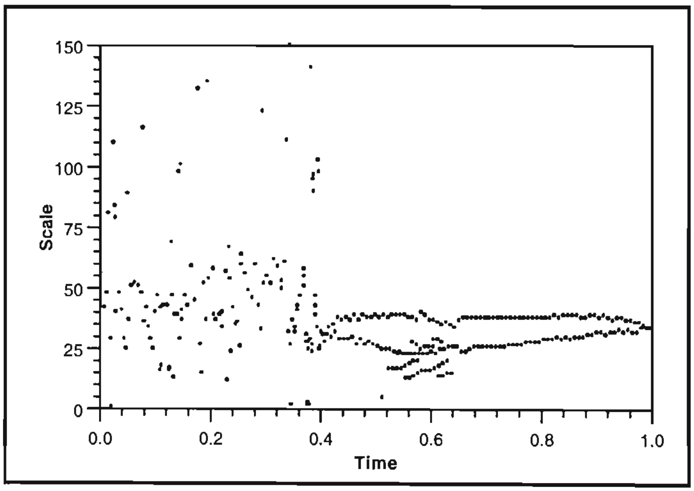


Figure 3-54: The time-frequency decomposition (DST) of AKC's spoken word: "three".

Figure 3-54 shows the Fourier coefficients superimposed on the SDST coefficients for the vowel /ee/. Both transforms produce two coefficients with two distinct spikes, albeit with slightly differing frequencies.

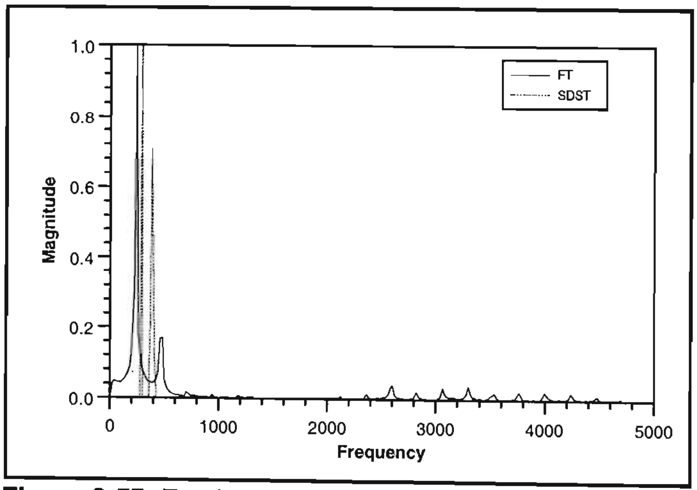


Figure 3-55: Fourier coefficients versus SDST coefficients for the vowel /ee/ of AKC's spoken word: "three".

DST and SDST function plots of "four"

Figure 3-55 shows the DST of AKC's spoken word "four". The start of the word, the fricative /f/, is associated with the usual randomness of the DST coefficients. Thereafter, the time-frequency plot of Figure 3-56 leaves no doubt that three distinct formants are used in the construction of the /aw/ vowel. Although the frequencies of these formants stay relatively constant, slight frequency changes occur. These changes exhibit interesting trends, with each formant's frequency seemingly independent of the others. Note the fall and rise in frequency of the fundamental, and slow decay of the higher frequency formants.

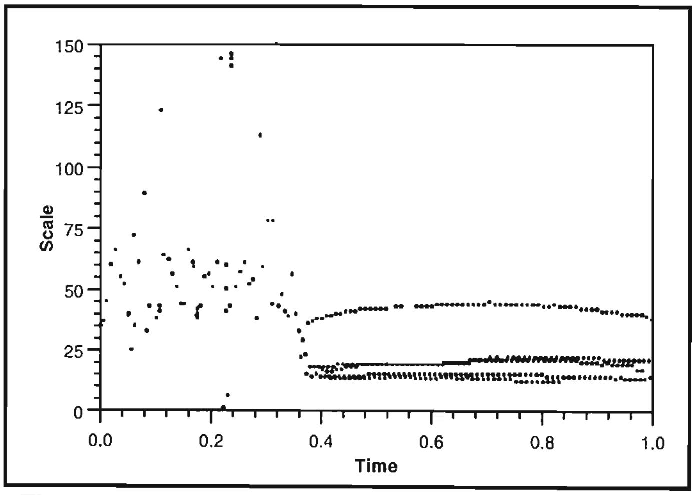


Figure 3-56: The time-frequency decomposition (DST) of AKC's spoken word: "four".

Results obtained with the FT and the SDST are very similar. The fundamental frequency from both transforms is equal. The high formants frequencies are slightly over-estimated. Note the transform's corresponding shape of the 'double-formant' around 800 Hz.

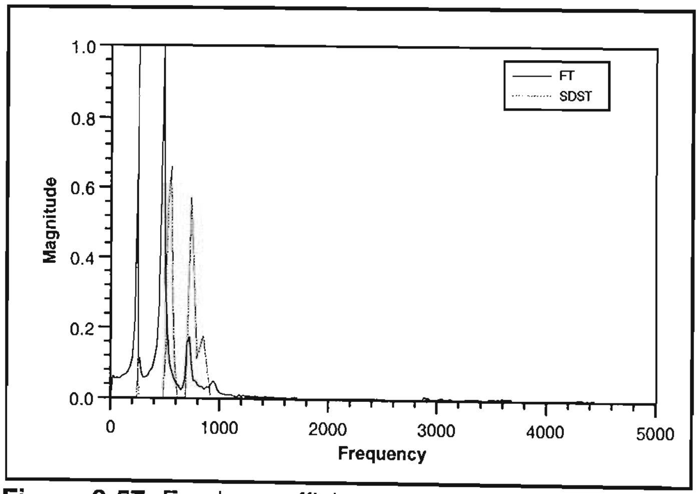


Figure 3-57: Fourier coefficients versus SDST coefficients for the vowel /aw/ of AKC's spoken word: "four".

DST and SDST function plots of "five"

Figure 3-57 shows the DST of AKC's spoken word "five". The voiceless fricative /f/ and voiced fricative /v/ at the start and end of the word are distinguishable. The formants in the vowel seem fairly constant except for a sudden change at around $t = 0.64$, where the number of formants appears to drop to three.

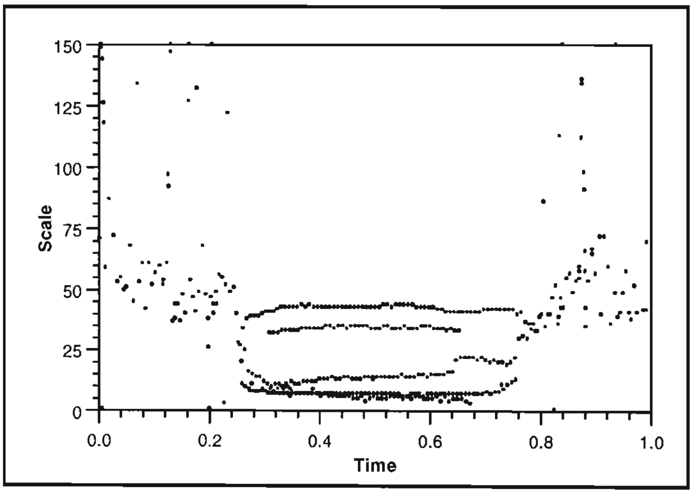


Figure 3-58: The time-frequency decomposition of AKC's spoken word: "five".

The SDST plot clearly shows the four apparent formants, whereas the FT has many aliasing spikes and no obvious formants. Once again, both transforms have accurately detected the fundamental.

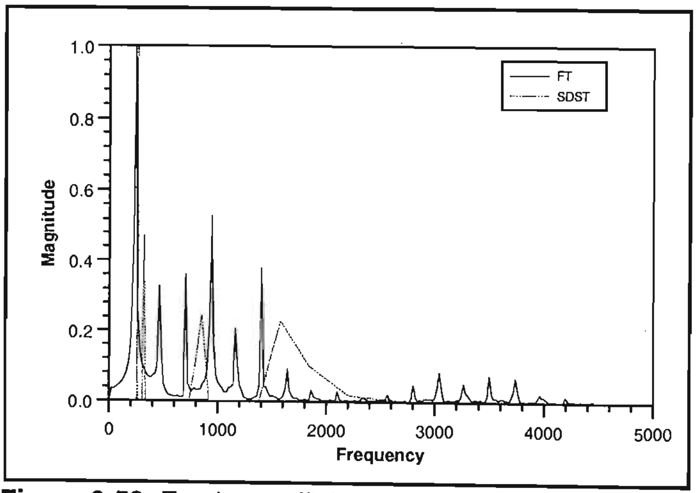


Figure 3-59: Fourier coefficients versus SDST coefficients for the vowel /i/ of AKC's spoken word: "five".

DST and SDST function plots of "six"

Figure 3-59 shows the DST of AKC's spoken word "six". The DST of the word "six" consists primarily of relatively random coefficients, due to the fricatives /s/ and /ks/. The /I/ fricative lasts only for a short time, and its formants are quite obvious from the DST.

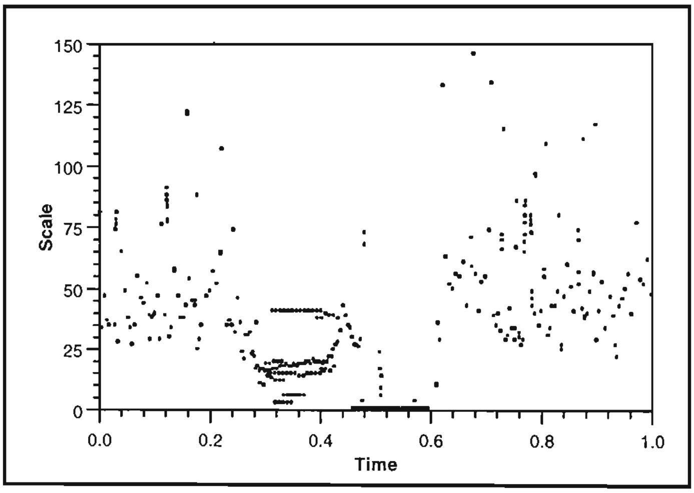


Figure 3-60: The time-frequency decomposition (DST) of AKC's spoken word: "six".

The three lower frequency formants are shown below corresponding to the FT and SDST. The general shapes of the graphs are exceptionally similar. Note the detection of the high frequency formant around 1800 Hz.

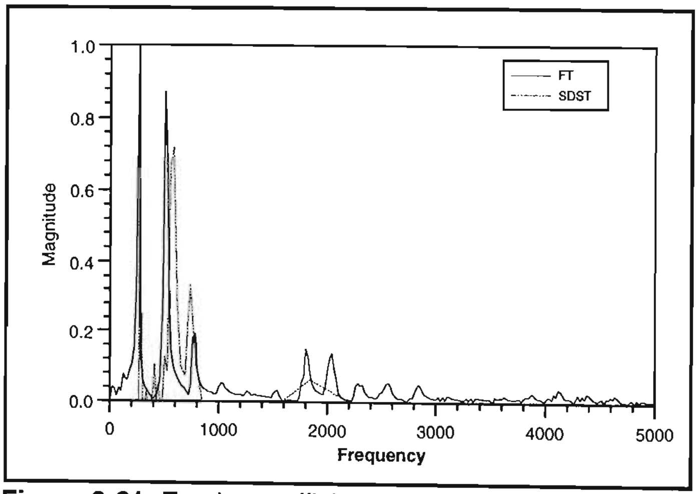


Figure 3-61: Fourier coefficients versus SDST coefficients for the vowel /i/ of AKC's spoken word: "six".

DST and SDST function plots of "seven"

Figure 3-61 shows the DST of AKC's spoken word "seven". The fricative /s/ begins the word with a very high frequency component represented here by small scale coefficients. This is followed by the vowel /ε/. The voiced fricative /v/ does not generate high WR/STD values usually associated with fricatives, as it is pronounced differently. The nasal /n/, for $t = [0.76..1]$, has a very characteristic DST shape. This is identical to its shape in the words "one" and "nine".

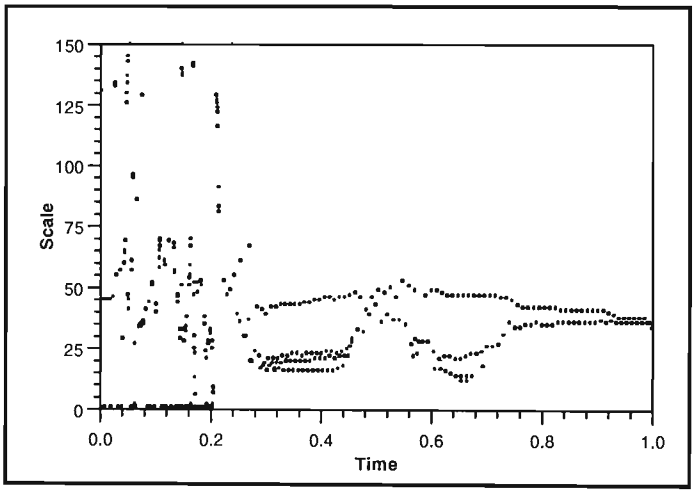


Figure 3-62: The time-frequency decomposition (DST) of AKC's spoken word: "seven".

The FT and SDST coefficients are shown below. These transforms represent the vowel /ε/. The three formants detected by both transforms have equal frequency.

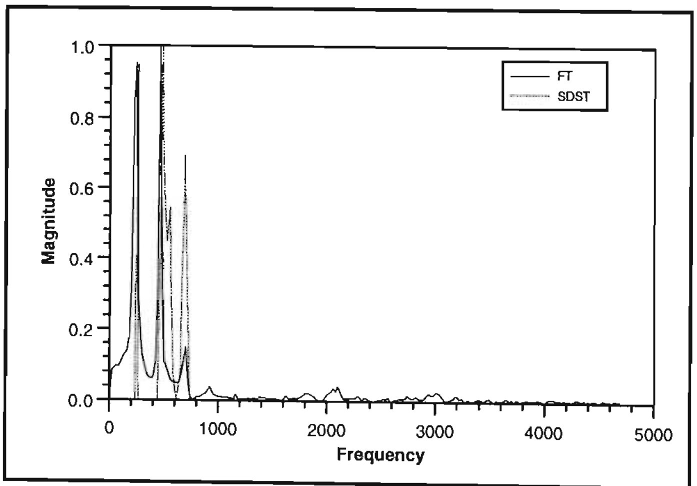


Figure 3-63: Fourier coefficients versus SDST coefficients for the vowel /ε/ of AKC's spoken word: "seven".

DST and SDST function plots of "eight"

Figure 3-63 shows the DST of AKC's spoken word "eight". The word "eight" is quite distinctive from the rest of the words in our set for it is the only one to begin with a vowel and end with a fricative. A stop separates the vowel and fricative. The DST and FT coefficients below isolate the formant frequencies as well as indicating the presence of a high frequency component. Note the very high frequency component associated with the fricative.

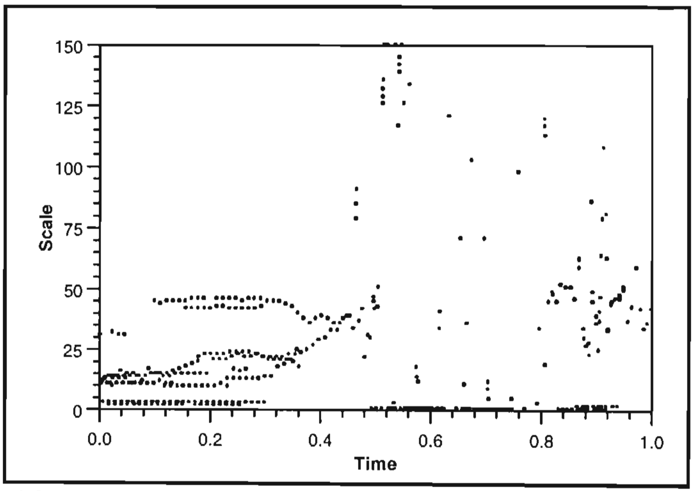


Figure 3-64: The time-frequency decomposition (DST) of AKC's spoken word: "eight".

Aliasing is again present in the FT plot. The high frequency component is represented by only a single non-zero coefficient at 3675 Hz. The fricative /t/ follows the stop from about $t = 0.8$.

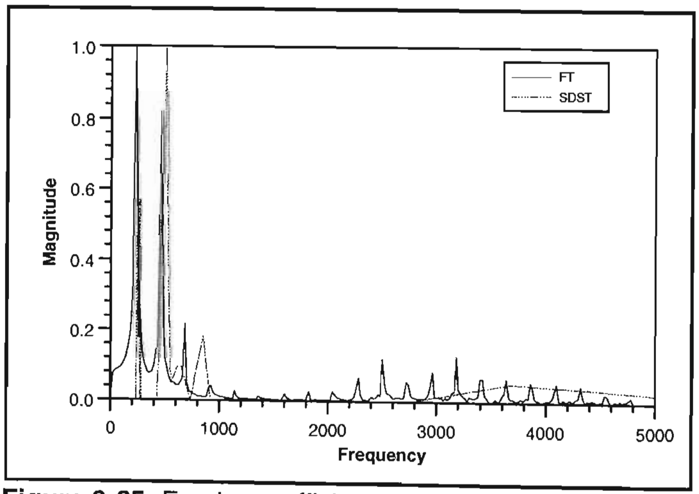


Figure 3-65: Fourier coefficients versus SDST coefficients for the vowel /ai/ of AKC's spoken word: "eight".

DST and SDST function plots of "nine"

Figure 3-65 shows the DST of AKC's spoken word "nine". The word begins and ends with a nasal /n/ with the vowel /ie/ occurring in-between. Two characteristic features to note are: firstly, the high frequency content occurring during the vowel and, secondly, the by now familiar shape of the terminating nasal /n/.

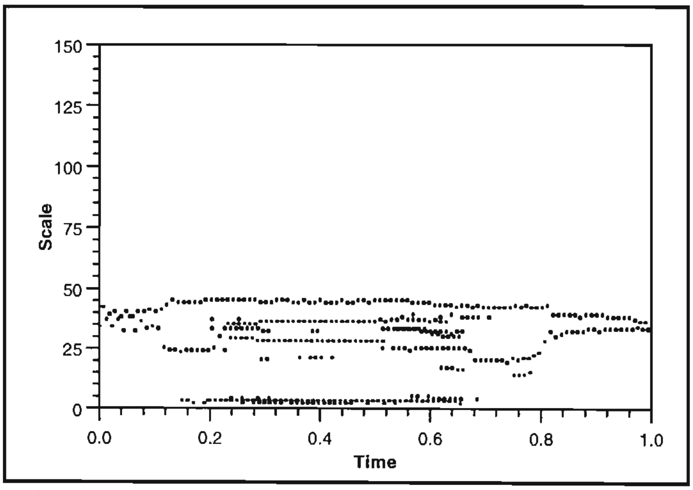


Figure 3-66: The time-frequency decomposition (DST) of AKC's spoken word: "nine".

Although the fundamental of both transforms concur, the coefficients of the formants do not match exactly. The aliasing is again clearly apparent in the FT. The detection of the high frequency component by the SDST is depicted by the two adjacent coefficients at the frequencies 3675 Hz and 5512 Hz.

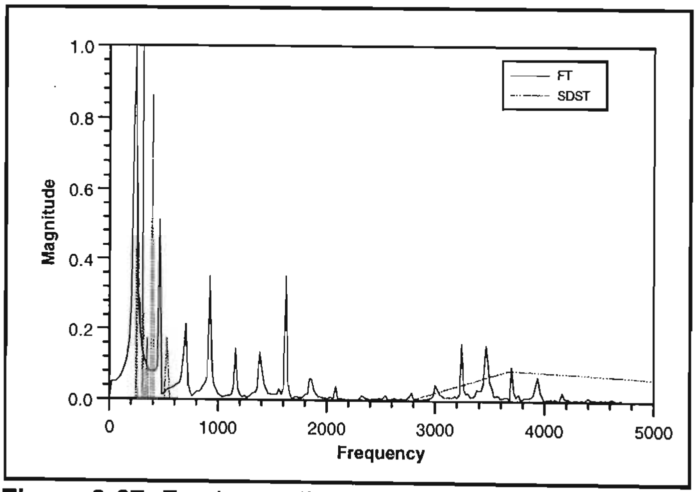


Figure 3-67: Fourier coefficients versus SDST coefficients for the vowel /ie/ of AKC's spoken word: "nine".

3.6 Speech Recognition Results

Digitally-recorded spoken words from each of the three speakers constitute our training and testing data sets. Stop, vowel, and fricative segmentation were responsible for the majority of the classification process. The separation into classes of the words comprises of the following binary decision structure.

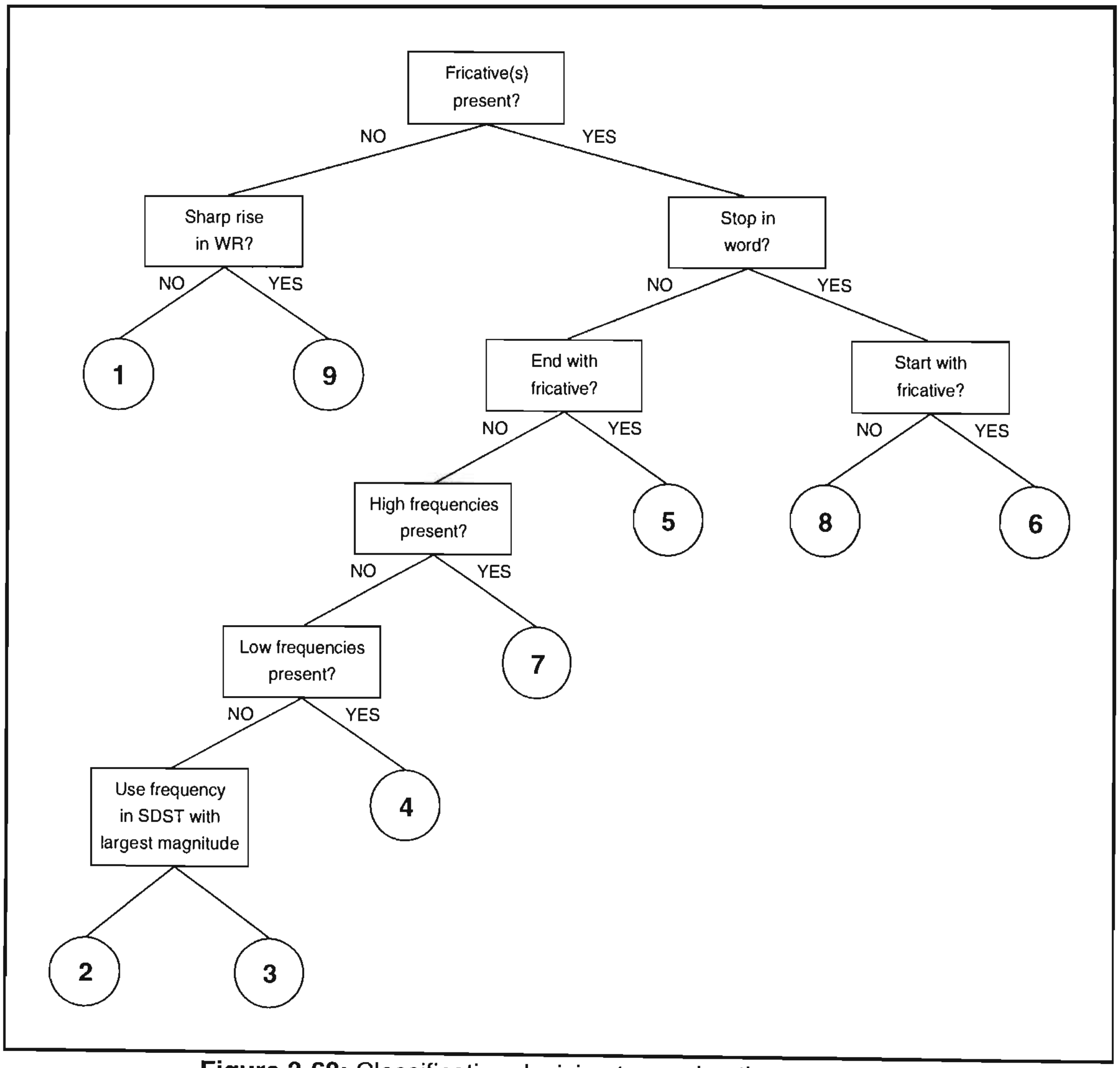


Figure 3-68: Classification decision tree using the presence of stops, vowels, and fricatives.

At each conditional stage in the binary tree, a decision is made based on the existence or non-existence of certain characteristics in the signal. The decision process relies on the following characteristics:

- **Fricatives:** As well as being speaker-independent, fricatives proved to be very easy to locate by the simple process of setting a suitable threshold level above which the WR values indicate a frication. The process obviously relies heavily on the existence of high frequencies in the frication regions. The low-cost microphone that we used attenuated these frequencies slightly. This resulted in two fricatives 'being missed'.

- **Stops:** Stops proved quite simple to detect. Clearly, this feature is the 'ultimate' in speaker-independence for there is no sound during a stop.

- **Vowels:** The detection of vowels is achieved by the SDST coefficients. In fact, only the coefficient having greatest magnitude was considered. The SDST coefficients were only examined once we knew, with a good deal of confidence, that the word was either a "two", "three", or "four". The word "four" has significantly larger dominant scales in the vowel region than do the other two words, and we achieved a 100% reliability in detecting the word "four". The author felt that the accuracy in correctly choosing between the words "two" and "three", although already quite high, could have been increased using, for example, Artificial Neural Networks with the SDST coefficients as the inputs.

Using a small training set of a few digits from each speaker, we estimated thresholds which would segment the words into satisfactory classes. A total of 90 words were processed for each speaker and the results appear in Tables 3-6, 3-7, and 3-8. Each of the digits, from each speaker, was tested 10 times and the reported prediction appears in the tables. The accuracy with which each digit was predicted is shown on the right of the table.

One male speaker (HG) and the one female speaker (AKC) achieved accurate recognition rates of just over 91%, whereas the speaker DIC achieved a success rate of just under 96%.

Speaker AKC											
	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**Accuracy**
1	1	1	1	1	1	1	1	1	1	1	100%
2	3	2	3	2	2	2	2	2	2	3	70%
3	3	3	2	3	3	2	2	2	2	3	50%
4	4	4	4	4	4	4	4	4	4	4	100%
5	5	5	5	5	5	5	5	5	5	5	100%
6	6	6	6	6	6	6	6	6	6	6	100%
7	7	7	7	7	7	7	7	7	7	7	100%
8	8	8	8	8	8	8	8	8	8	8	100%
9	9	9	9	9	9	9	9	9	9	9	100%
Total accuracy over the entire set											91%

Table 3-6: Speech recognition accuracy for the speaker AKC.

The digits containing stops and fricatives were all reported with 100% accuracy. Our crude and simplistic coefficient analyser, which simply examines the fundamental, often does not have sufficient information to decide between the digit "two" and the digit "three". We therefore propose that a more advanced pattern recognition system be used, i.e., one which takes all the formants into consideration. We suggest Artificial Neural Networks as a possible solution.

Speaker DIC											
	A	B	C	D	E	F	G	H	I	J	Accuracy
1	1	1	1	1	1	1	1	1	1	1	100%
2	2	2	2	3	2	2	2	2	2	2	90%
3	2	3	3	2	3	3	3	3	3	2	70%
4	4	4	4	4	4	4	4	4	4	4	100%
5	5	5	5	5	5	5	5	5	5	5	100%
6	6	6	6	6	6	6	6	6	6	6	100%
7	7	7	7	7	7	7	7	7	7	7	100%
8	8	8	8	8	8	8	8	8	8	8	100%
9	9	9	9	9	9	9	9	9	9	9	100%
Total accuracy over the entire set											96%

Table 3-7: Speech recognition accuracy for the speaker DIC.

Speaker HG											
	A	B	C	D	E	F	G	H	I	J	Accuracy
1	1	1	1	1	1	1	1	1	1	1	100%
2	2	2	2	2	2	9	2	2	2	1	80%
3	3	3	3	3	3	3	3	6	3	2	80%
4	4	4	4	4	4	4	4	4	4	4	100%
5	5	5	5	5	5	5	5	5	5	5	100%
6	6	6	6	6	6	6	6	6	6	6	100%
7	7	7	7	7	7	7	7	7	7	7	100%
8	8	9	9	8	8	8	8	8	8	8	80%
9	9	9	9	9	9	9	9	9	9	9	100%
Total accuracy over the entire set											91%

Table 3-8: Speech recognition accuracy for the speaker HG.

3.7 FUTURE RESEARCH POTENTIAL

- *Speaker independence* is a characteristic supported by a speech recognition system that operates without regard for which particular individual uses the system. The more speaker-independent the system is made, the smaller the difference in performance between various speakers will be. There are many variables in this scenario, including speaker sex, nationality, age, and even whether the speaker is a smoker or a non-smoker. Trying to achieve speaker-independence is a difficult task, but one made easier by the maximum utilisation of speaker-independent features, as we have tried with the location of fricatives. Clearly, however, there is no easy way of circumventing the problems associated with formant frequency identification.

- *Artificial Neural Networks* (ANN) have emerged as a powerful non-linear classifier and has performed well on speech classification [Low89]. Our crude vowel recognition module bases its answer entirely on the fundamental. Intuition tells us that any word recognition system which does base its vowel determination entirely on the fundamental whilst ignoring the remaining formants should not expect very good results. A preferred method would be to interpret each of the reported formants. The SDST provides a set of coefficients which are directly usable for the inputs to the input layer neurons of an ANN since the pre-processing of the signal has been accomplished and each of the coefficients relate to a specific characteristic of the signal. In this regard, the author has no doubt that the use of ANN would have led to slightly better results.

- *Larger vocabularies* tend to require more computation, storage, and time-consuming training sessions. The response time and error rates of the system also tend to increase linearly with an increase in vocabulary size.

- *Continuous speech recognition systems* are much more difficult to design than isolated speech recognition systems. Throughout our research, each signal was manually broken up into separate signals containing only the samples pertaining to that word. These signals were each dealt with individually. If a continuous speech recognition system is required, the system has to segment particular characteristics which are then built up throughout the signal. Word boundaries must then be decided upon using this information. The removal of information (word boundaries) obviously tends to increase the error rate.

- *Time-warping techniques* [Vai85] eliminate most of the problems associated with speaking rate. This attribute is partly related to speaker-independence although not entirely. Different speakers will, in general, speak at different rates. We can, in some circumstances, use time-warping to our benefit. For example, words like "bid" and "bit" differ mainly in the duration of the vowel.

- *Speaker recognition* is currently an extremely important topic in the commercial arena. For example, Automatic Teller Machines (ATMs) currently 'know' which card they are issuing money to, but not which person, hence the tremendous fraud rate concerning ATMs [Cas91]. Clearly, it is in the best interests of the organisation to know to which individual they are dealing with. Speaker recognition over telephone lines runs into the usual bandlimitedness of the currently installed lines. This makes the task significantly more difficult. Experimentation has been conducted by other researchers into 'cough recognition', whereby a signature of the vocal tract is registered and stored as a template for future matching. The time-frequency space of the DST could provide sufficient information upon which to base a decision.

3.8 Chapter Summary

This chapter has shown that the Dominant Scale Transform performs exceptionally well in the decomposing of signals in the time domain into a time-frequency space. When applied to speech, not only is the fundamental represented, but so too are several of the formants, making vowel recognition a simple pattern recognition problem. The DST was also used for the classification of silence regions, regions of voiced speech, and regions of frication. The broad frequency spectrum exhibited in a region of frication translates into rapidly changing DST coefficients which would be roughly equivalent to a Fourier Transform of the fricative. This output proved very reliable and perfectly consistent with other authors in the field who used completely different methods. Further work is required to define a model whereby the several different fricatives can be identified.

An extremely powerful result was illustrated, namely, the summation of an interval of DST coefficients, and hence the SDST coefficients, approximate the FT of that interval remarkably closely. This opens the door to possible replacement of the FT by the SDST in some select applications.

We concluded the chapter by presenting results obtained from our DST/SDST based speech recognition system. Although a very simplistic formant recogniser was used, our speech recognition system obtained very high recognition rates.

Chapter 4

Using the DST for the Time-Frequency decomposition of various Real-World Applications

4.1 INTRODUCTION

It is unrealistic to assume the signals occurring in real-life are periodic and have no frequency transients. In fact, the opposite is probably more accurate. Imagine listening to a pure sine wave for the whole day! The melody, or a sequence of notes, communicates much of the beauty in music. We now focus our attention to three real-world applications which benefit from the spatially localised coefficients of the Dominant Scale Transform in the time-frequency space.

4.2 THE CREATION OF A MUSIC SCORE FROM A MUSIC SIGNAL

Music signals are rich in both harmonics and frequency transients. The efficiency of the DST makes it very suitable for the economical incorporation into a range of hardware apparatus. For example, a hand-held music score generator is well within the realms of implementation. A musical score indicates which notes have to be played at consecutive time steps. It is, therefore, a time-frequency analysis which is much closer to a Short-Time Fourier Transform than to the Fourier Transform, where all association of time is lost, or at least not explicitly recognisable. The pitch of a note played on any instrument is determined by the fundamental frequency. Additional harmonics and overtones alter the shape of the wave to give each instrument its characteristic sound.

The keyboard (see Appendix A) was set to mimic a flute and an organ. The A note above middle C was played and the waveforms appear in Figure 4-1. Surprisingly, the fundamental frequency of the organ was exactly one octave below that of the flute, which gives the organ its distinctive deep sound. The DST coefficients obtained from these signals are shown in Figures 4-2 and 4-3. Note the detection of harmonics in Figure 4-3.

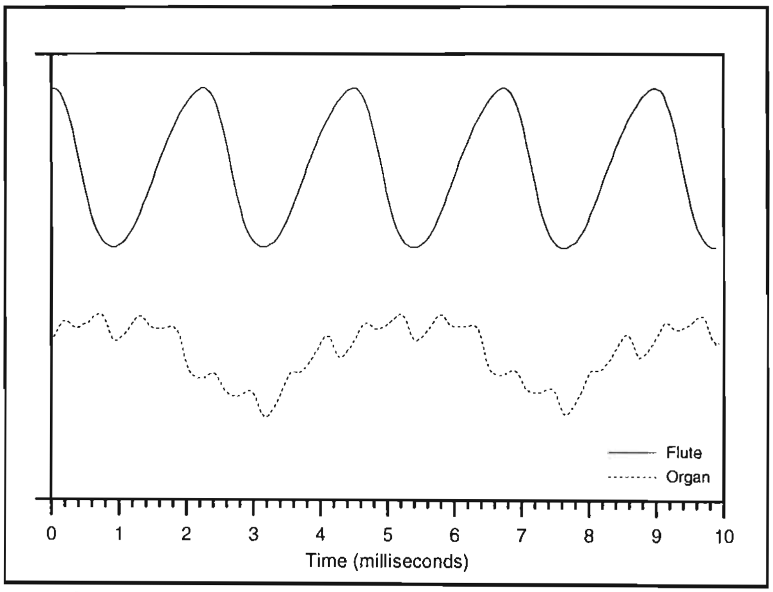


Figure 4-1: The flute and organ signals of an A note.

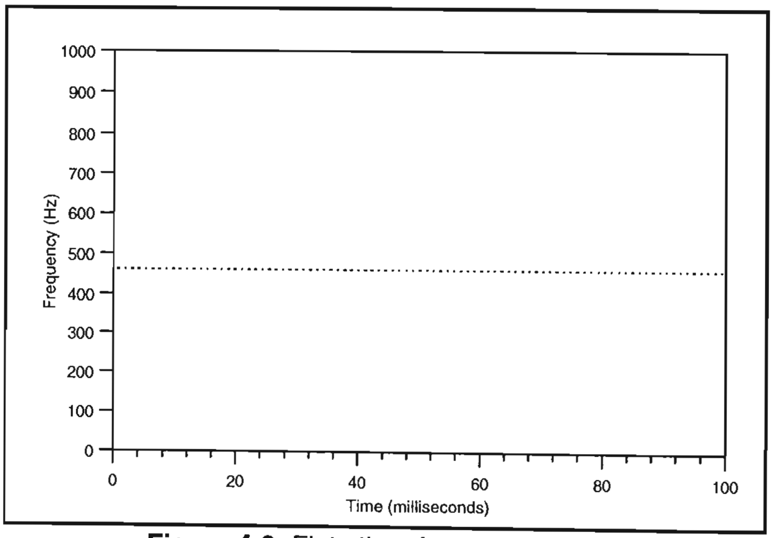


Figure 4-2: Flute time-frequency space.

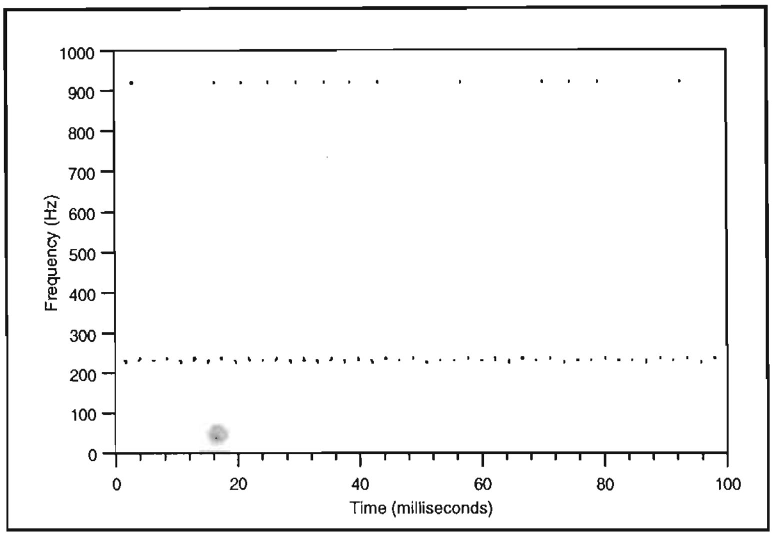


Figure 4-3: Organ time-frequency space.

The fundamental frequency of an A note is exactly 440 Hz. The undersampling of the DST is once again evident by the overrating of the frequency. The recorded fundamental frequency for the flute and organ is 230 Hz and 459 Hz respectively. The last example consists of the first nine notes of the well-know Beethoven piece Fur Elise played (albeit a bit fast!) by the author using the flute setting. The time-frequency space has localised each note accurately. Note that for the sake of display purposes only, a suitable threshold was set, below which no coefficients were displayed.

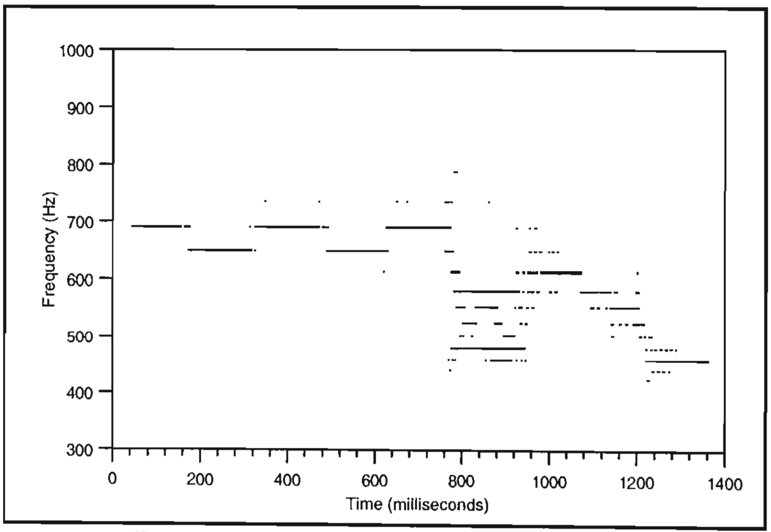


Figure 4-4: Fur Elise time-frequency space.

4.3 Projectile Spin Behaviour

Telemetry data from a rotating object was received as a data series consisting of approximately 20,000 points sampled with the time interval Δt = 3 ms. This problem was initially tackled by Prof. M.J. Alport and part of his work appears here. The object's rotation accelerates from about 0 Hz to about 20 Hz and then slowly decays.

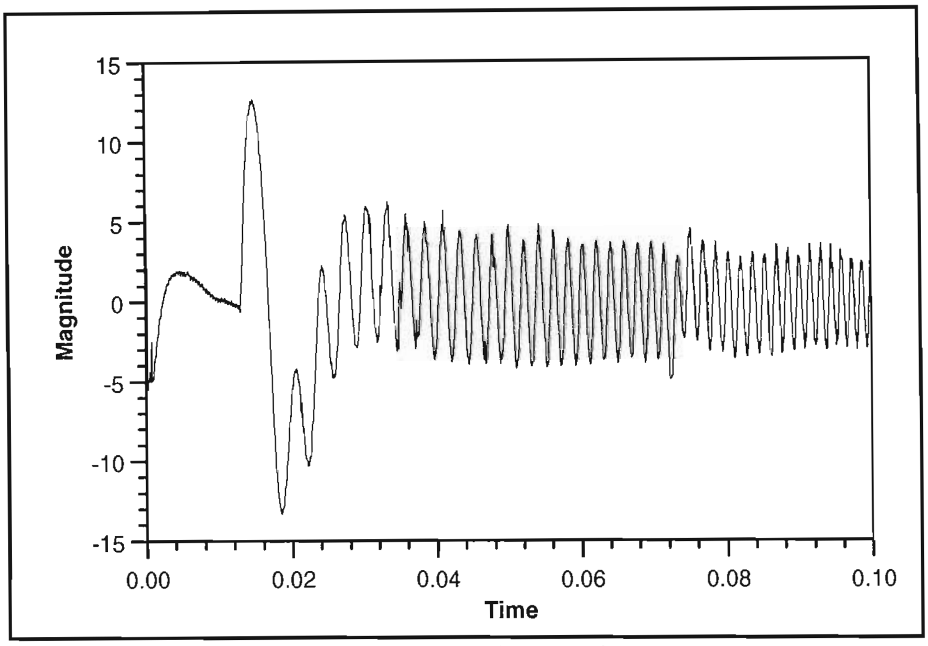


Figure 4-5: The start of the sample data set.

4.3.1 The Short-Time Fourier Transform approach

The following STFT work was performed by Prof. M.J. Alport and his method and description appears below. Usually, FFT techniques can only be applied to a time series of data points that has a stationary spectrum. Clearly, the present time series does not satisfy this criterion. Although sophisticated techniques such as the Gabor Transform may be employed to analyse such non-stationary data, in the present case, since the sampling frequency is a good factor of 10 higher than the average spin frequency, a simpler technique is used. This simpler technique involves decimating the time series into a number of 512 byte sub-samples and then performing an FFT analysis on these sequential sets. To improve the temporal response of the analysis, the 512 byte samples are overlapped by 256 bytes with their neighbours. Within each 512 byte sample, the magnitude of the strongest frequency component is determined using a simple peak search method. Fortunately, once the DC has been suppressed, the spin frequency is the dominant component. This technique gives a temporal resolution of:

$$\frac{17000 \times 0.003}{70} = 0.7 \text{ seconds} \qquad (4\text{-}1)$$

The data covers a total time period of approximately 50 seconds. To reduce the quantisation effects in the frequency analysis, the signal frequency is obtained by a parabolic interpolation about the central maximum. In addition, the data is tapered using a Hanning window. The peak spin frequency is also obtained by curve-fitting a second-order polynomial.

Figures 4-6 show the time versus frequency plot using Short-Time Fourier Transforms. The plot shows a very rapid increase in rotational velocity reaching a peak at around 10 seconds and then slowly decaying till around 40 seconds and then (for reasons unknown to the author) appearing to speed up! At about 4 seconds, there is a small 'kink' in the graph. It appears that the acceleration tapers off, and then suddenly increases again. The resolution of the windowing procedure applied to analyse the data using the STFT is not sufficiently small to gather much knowledge about the behaviour in that critical region.

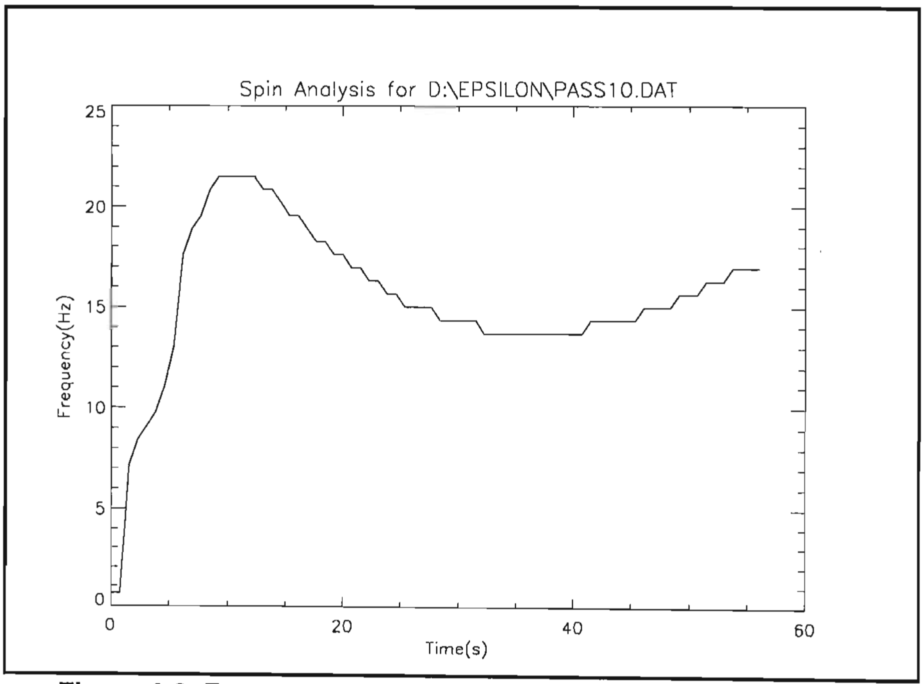


Figure 4-6: Frequency versus time for sample data set using STFT.

4.3.2 The Dominant Scale Transform approach

Using the same data series, we now present the results from the Dominant Scale Transform. Please note that we have normalised both the time and frequency axes. Also, all coefficients have been plotted. If a more defined plot was required, some threshold could be included whereby only coefficients having magnitude greater than the threshold would be plotted. The trend is almost exactly the same as that determined by the series of STFT calculations.

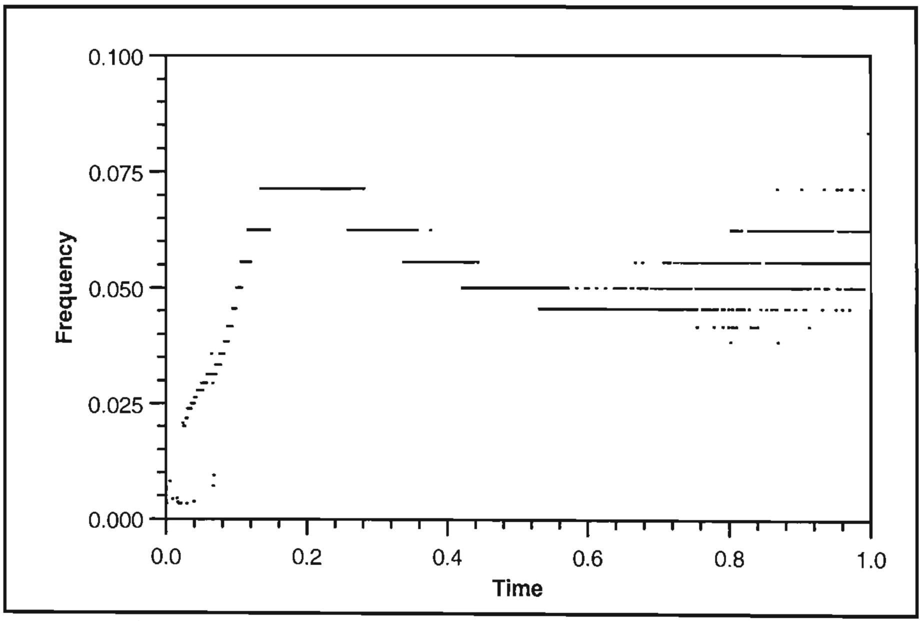


Figure 4-7: Time-frequency space of the sample data set.

There exists a short-lived phenomenon, occurring at $t = [0.06..0.07]$, which does in fact go undetected by the Fourier analysis method. The spectral bleeding between windows tends to blur this region, making it appear that the rotational acceleration decreases and then increases again all the time staying largely positive. The spatial localisation properties of the DST show that this is not, in fact, the case.

Careful examination of the time-frequency space of the DST, reveals that, during $t = [0.06..0.07]$, the rotational acceleration appears to stop, i.e. $\dot{\omega} = 0$ where ω is the rotational velocity in this case. The detection of this very short-lived phenomenon emphasises the importance of the accurate spatialisation provided by the DST.

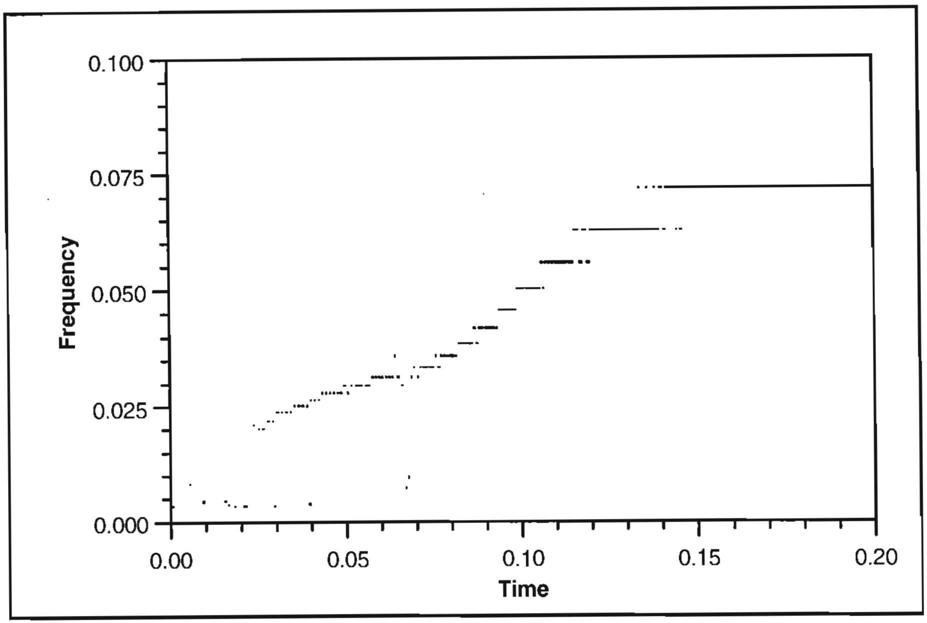


Figure 4-8: Portion of the time-frequency space of the sample data set.

4.4 Periodic Pulling

Very late into the author's research, Prof. M. J. Alport suggested that I use the DST to verify some results pertaining to periodic pulling [Las69] which were documented in [Koe93]. We refer the reader to [Koe93] for all the details pertaining to the experiment.

Very careful examination shows that there certainly appears to be a correlation between the frequency and magnitude of the signal. The reader should look at the trend along the bottom of the 'dots' in Figure 4-10. The scales can be seen to increase slightly in sympathy with the increased magnitude of the plasma density oscillations. Although this is very preliminary data, the author is encouraged by the output. This result appears to be in agreement with the result obtained in [Koe93].

Note: A 16:1 interpolation was effected on the original signal to improve the frequency resolution in the Scale Function. This interpolation shifts the scale of the fundamental frequency into the frequency range in which the DST scales are more dense - thereby increasing the scale (and hence frequency) resolution.

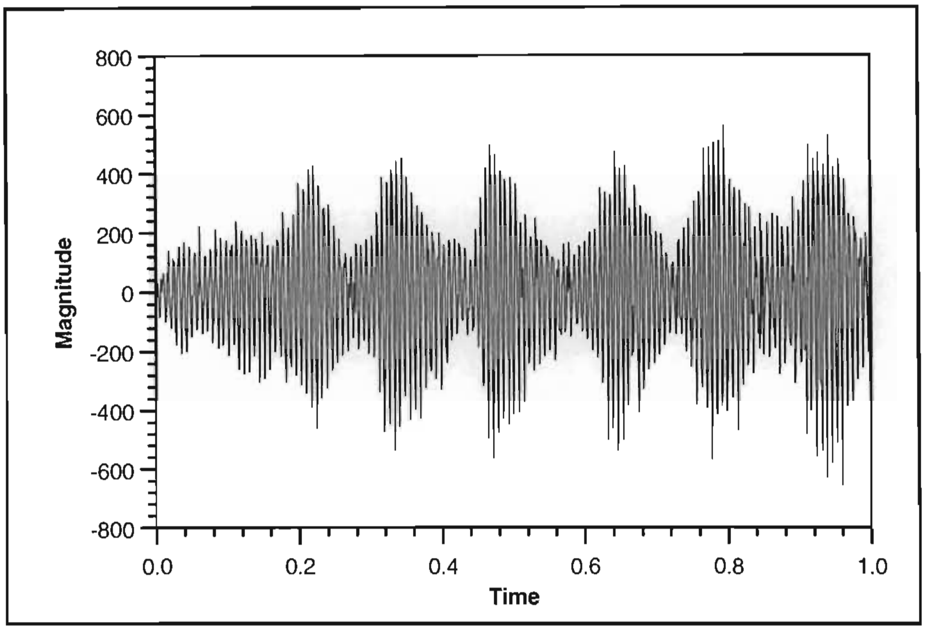


Figure 4-9: Plasma density oscillations.

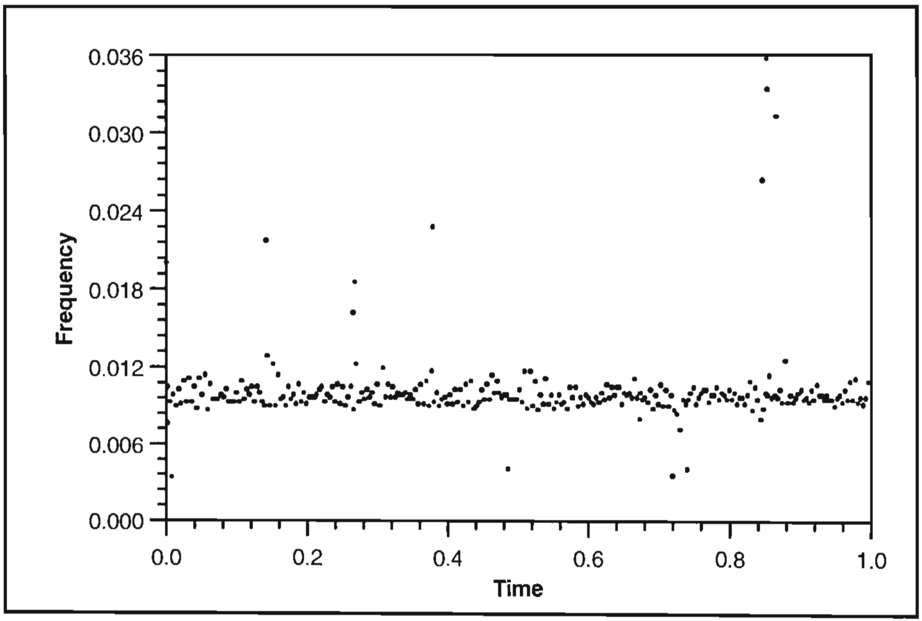


Figure 4-10: Modulation of the instantaneous frequency.

Chapter 5

Conclusion

The thesis began by examining several situations in which the well-established Fourier Transform's performance was less than desired. These examples contained frequency transients and singularities which have the potential to change the Fourier coefficients quite unexpectedly.

5.1 ACHIEVEMENTS OF THE DST

Chapter 2 introduced the concept of the wavelet and went on to describe the new Dominant Scale Algorithm and its associated Dominant Scale Transform. The algorithm is an entirely integer-based $O(n)$ algorithm which approximates a time-domain signal in its time-frequency space. We feel the Dominant Scale Transform makes a significant contribution to signal processing for the following reasons:

- *Frequency tracking* as a function of time is possible. The formant tracking shown in Chapter 3 provided excellent time-frequency space plots in which formant frequency can be seen to change with time.

- *Phase tracking* as a function of time can be very useful in many real-time applications. Any phase changes occurring in the signal, or even in a particular component, will be localised accurately in time. This attribute was discussed in Chapter 2.

- *Singularities* which last for small t can be spatially localised very accurately in time and effectively isolated from the remainder of the signal attributable to the local support of wavelets.

- *Zooming* is a property of wavelets which describes the altering of the region of support of the basis functions, enabling the detection and localisation of small perturbations. The use of wavelets as the basis functions in the DST means that the DST also exhibits this desired property.

- The Dominant Scale Algorithm is an *extremely efficient O(n) algorithm* which, due to its large parallelism nature, is ideal for implementation in hardware.

We have shown in Chapter 3 that, using the DST coefficients, a limited vocabulary speech recognition system can be implemented which exhibits extremely high recognition percentages. Moreover, the DST coefficients were shown to be entirely consistent with results obtained from other researchers in the field using entirely different methods. Formant frequencies closely matched those of the Fourier Transform and the author envisages this property allowing the DST to replace the Fourier Transform in selected applications. Three such applications were described in Chapter 4 and the advantages offered were described. These are just three applications in which the DST imparted more information than the equivalent Fourier could have, but there are obviously thousands which may benefit from the excellent time-frequency decomposition characteristics of the DST.

5.2 Future Research Opportunities Using the DST

The time-frequency space conveys spatial information which could be used in many more applications than described in this text. The possibilities are almost endless. We now describe some cases where the employment of the DST may provide the vital spatial information required to successfully achieve a given task. Some areas for future research are:

- The frequency difference between adjacent coefficients in the frequency scale of the DST tends asymptotically to zero. Most of the scales therefore represent the lower frequencies. *Reducing the number of coefficients* below a certain frequency would speed up the algorithm, however, the author is concerned that the removal of some of the scales may cause introduce inconsistencies and loss of stability.

- The author strongly suspects that the alteration of the *mother wavelet shape*, may change the characteristics and perhaps the accuracy of the time-frequency decomposition. Preliminary tests showed that no significant changes in the decomposition properties occurred. However, future research by the author is required to determine the extent to which the DST is affected by the wavelet shape.

- The research described in the thesis has concentrated entirely on the decomposition of the signal from the time domain into the time-frequency space. The author feels it is possible that approximate *signal reconstruction* could be achieved using the time-frequency space. Also, due to the sparseness, tremendous *compression ratios* should be achievable. These compression schemes may be very useful in the digital transmission of speech. By sending only the formant frequencies and some form of amplitude envelope, complex vowel sounds could be reconstructed at the receiver.

- The human speech production mechanism is unique to every individual. By producing sound which requires the use of all or most of the production system, we could, by using the DST time-frequency space, create a 'fingerprint' for any individual . A *speaker verification system* would use this fingerprint as a means of matching the spoken word with the individual speaking.

Bibliography

[Ade87] Adelson E.H. and Simoncelli E. (1987), "Orthogonal pyramid transforms for image coding", *Visual Communications and Image Processing*, Vol. 845

[Aus92] Auscher P. (1992), "Wavelet Bases for $L^2(\mathfrak{R})$ with Rational Dilation Factor", Wavelets and Their Applications, ISBN 0-86720-225-4, Pages 439-451

[Bad91] Badin P. (1991), "Fricative consonants: acoustic and X-ray measurements", *Journal of Phonetics*, Vol. 19, Pages 397-408

[Bey91] Beylkin G., Coifman R., & Rokhlin V. (1991), "Fast wavelet transforms and numerical algorithms", *Communications on Pure and Applied Mathematics*, ✓ Vol. 44, Pages 141-183

[Bra65] Bracewell R.M. (1965), The Fourier Transform and its Applications, McGraw-Hill (New York)

[Bra83] Bracewell R.M. (1983), "Discrete Hartley Transform", J. Opt. Soc. Am., Vol. 73, No. 12, Pages 1832-1835

[Bra84] Bracewell R.M. (1984), "The Hartley Transform", *Proceedings of IEEE*, Vol. 72, No. 8, Pages 1010-1018

[Bri84] Bristow G. (1984), "Electronic Speech Synthesis", Granada Publishing Ltd. (London), ISBN 0-246-11897-0

[Cas91] Cassidy G.A. (1991), "Voice Recognition", *Second South African Workshop on Pattern Recognition*, Pages 157-166

[Chu92a] Chui C.K. (1992), "An Introduction To Wavelets", Wavelet Analysis and its Applications, Vol. 1, Academic Press Inc., London, ISBN 0-12-174584-8

[Chu92b] Chui C.K. (1992), "An Tutorial in Theory and Applications", Wavelet Analysis and its Applications, Vol. 2, Academic Press Inc., London, ISBN 0-12-174590-2

[Chu92c] Chui C.K. (1992), "On Cardinal Spline-Wavelets", Wavelets and Their Applications, ISBN 0-86720-225-4, Pages 419-438

[Coo65] Cooley J.W., and Tukey J.W. (1965), "An algorithm for machine calculation of complex Fourier series", *Math. Computation*, Vol. 19, Pages 297-301

[Cun92] Cunningham E.P. (1992), "Digital Filtering: An Introduction", Houghton Mifflin Company (Boston), ISBN 0-395-53989-7

[Dau86] Daubechies I., Grossman A., and Meyer Y. (1986), "Painless non-orthogonal expansions", J. Math. Phys., Vol. 27, No. 5, Pages 1271-1283

[Dau88a] Daubechies I. (1988), "Orthonormal Bases of Compactly Supported Wavelets", *Communications on Pure and Applied Mathematics*, Vol. XLI, Pages 909-996

[Dau88b] Daubechies I. and Paul T. (1988), "Time-frequency localisation operators - a geometric phase space approach: II. The use of dilations", *Inverse Problems*, Vol. 4, Pages 661-680

[Dau90] Daubechies I. (1990), "The Wavelet Transform, Time-Frequency Localization and Signal Analysis", *IEEE Transactions on Information Theory*, Vol. 36, No. 5, Pages 961-1005 ✓

[Dau92] Daubechies I. (1992), "Ten Lectures on Wavelets", *CBMS-NSF Regional Conference Series in Applied Mathematics,* SIAM, ISBN 0-89871-274-2

[Dre72] Dreyfus-graf J. (1972), "Parole codée (phonocode): reconnaissance automatique de langages naturels et artificiels", *Revue d'Acoustique*, No. 21, Pages 3-12

[Fal85] Fallside F. and Woods W.A. (1985), "Computer Speech Processing", Prentice/Hall International, ISBN 0-13-163841-6

[Fla90] Flandrin P., Magand F., and Zakharia M. (1990), "Generalized Target Description and Wavelet Decomposition", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 2, Page 350-352 ✓

[Fla72] Flanagan J.L. (1972), "Speech Analyis, Synthesis, and Perception", Springer-Verlag (New York)

[Gab46] Gabor D. (1946), "Theory of communication", *Journal Inst. Elect. Eng.*, Vol. 93, No. III, Pages 429-457

[Gol93a] Goldstein H. (1993), "A practical approach to using Wavelets in Signal Processing: The Dominant Scale Transform", *Proceedings of the Nineteenth South African Symposium on Numerical Mathematics*, Pages 67-76 ✓

[Gol93b] Goldstein H. (1993), "On Dominant Scales of 1D signals using a variant Haar Wavelet", *South African Institute Of Electrical Engineers*, Vol. 84, No. 3, Pages 218-224, September 1993

[Gol93c] Goldstein H. (1993), "Using non-orthogonal wavelets in the spatial and spectral decomposition of music signals", *Fourth South African Workshop on Pattern Recognition*, Pages 87-92, November 1993

[Gol94] Goldstein H., "Formant tracking using the wavelet-based DST", *IEEE COMSIG-94*, Pages 183-189, October 1994, ISBN 0-7803-1998-2

[Gou84] Goupillaud P. (1984/5), Grossman A., and Morlet J., "Cycle-Octave and Related Transforms in Seismic Signal Analysis", *Geoexploration*, Vol. 23

[Gre78] Greene M.C.L. (1978), "The Voice and its Disorders", *Pitman Medical*, London, Pages 1-45

[Gro84] Grossman A. and Morlet J. (1984), "Decomposition of Hardy Function into Square Integrable Wavelets of Constant Shape", SIAM J. Math. Anal., Vol. 15, No. 4

[Ham83] Hamming R.W. (1983), "Digital Filters", Second Edition, ISBN 0-13-212506-4, Prentice-Hall Inc.

[Hei61] Heinz J.M. and Stevens K.N. (1961), "On the Properties of Voiceless Fricative Consonants", *The Journal of the Acoustical Society of America*, Vol. 33, No. 5

[Hei89] Heil C.E. and Walnut D.F. (1989), "Continuous and Discrete Wavelet Transforms", *SIAM Review*, Vol. 31, No. 4 ✓

[Hol88] Holmes J.N. (1988), "Speech Synthesis and Recognition - Aspects of Information Technology", Van Nostrand Reinhold (UK) Company Ltd., ISBN 0-278-00013-4

[Hug56] Hughes G.W. and Halle M. (1956), "Spectral Properties of Fricative Consonants", *The Journal of the Acoustical Society of America*, Vol. 28, No. 2

[Kno90] Knowles G. (1990), "VLSI Architecture for the Discrete Wavelet Transform", *Electronic Letters*, Vol. 26, No. 5, Pages 1184-1185

[Koe93] Koepke M.E., Alport M.J., Sheridan T.E., Amatucci W.E., and Carrol J.J. III (1993), "Asymmetric Spectral Broadening of Driven Electrostatic Ion-Cyclotron Waves", submitted for publication.

[Kro87] Kronland-Martinet R., Morlet J. and Grossman A. (1987), "Analysis of Sound Patterns Through Wavelet Transforms", *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 1, No. 2, Pages 273-301

[Lad73] Ladefoged P. (1973), "Elements of Acoustic Phonetics", The University Of Chicago Press

[Lad85] Ladefoged P. (1985), "The Phonetic Basis for Computer Speech Processing", Computer Speech Processing, Edited by Frank Fall side and William A. Woods, ISBN 0-13-163841-6, Prentice-Hall International (UK) Ltd.

[Lan88] Lancaster D. (1988), "Active-Filter Cookbook", Howard W. Sams and Company, ISBN 0-672-21168-8

[Las69] Lashinsky H. (1969), "Periodic pulling and the transition to turbulence in a system with discrete modes", *Turbulence of Fluids and Plasmas*, edited by J. Fox, page 29, Polytechnic, Brooklyn (New York).

[Lee88] Lee, E.A., and Messerschmitt D.G. (1988), "Digital Communication", Kluwer Academic Publishers, ISBN 0-89838-274-2, Pages 188-206

[Lew89] Lewis A.S. and Knowles G. (1989), "Video Compression using 3D Wavelet Transforms", *Electronic Letters*, Vol. 26, No. 6, Pages 396-398

[Lie61] Lieberman P. (1961), "Pertubations in Vocal Pitch", *The Journal of the Acoustical Society of America*, Vol. 33, No. 5

[Low89] Lowe D. (1989), "Adaptive radial basis function nonlinearities and the problem of generalisation", *Proceedings of IEE Conference on Artificial Neural Networks*

[Mal89a] Mallat S.G. (1989), "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 7 ✓

[Mal89b] Mallat S.G. (1989), "Multiresolution Approximations and Wavelet Orthogonal Bases of $L^2(\Re)$", *Transactions of the American Mathematical Society*, Vol. 315, No. 1 ✓

[Mal89c] Mallat S.G. (1989), "Multi-frequency Channel Decompositions of Images and Wavelet Models", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 12

[Mal92] Mallat S.G. and Hwang W.L. (1992), "Singularity Detection and Processing with Wavelets", *IEEE Transactions on Information Theory*, Vol. 38, No. 2, Pages 617-643 ✓

[Mey86] Meyer Y. (1986), "Ondelletes, Fonctions splines et analyses graduées, Lectures given at the University of Torino, Italy.

[Mey90] Meyer Y. (1990), *Ondellettes*, Hermann

[Mey93] Meyer K. (1993), "The Fast Wavelet Transform and Recognising Spoken Numbers", *Fourth South African Workshop on Pattern Recognition*, Pages 196-204, November 1993

[Mil85] Millar P.C. (1985), "Recursive Quadrature Mirror Filters - Criteria Specification and Design Method", *IEEE Transaction of Acoustics, Speech, and Signal Processing*, Vol. ASSP-33, No. 2

[Mor82] Morlet J., Arens G., Fourgeaui., and Giard D. (1982), "Wave Propagation and Sampling Theory", *Geophysics*, Vol. 47, Pages 203-236

[Mor83] Morlet J. (1983), "Sampling Theory and Wave Propagation", in *NATO ASI Series*, Vol. 1, Issues in Acoustic signal/Image processing and recognition, Chen C. H., ed., Springer-Verlag, Berlin, Pages 233-261

[Mur90] Murrell H.C. and Carson D.I. (1990), "Image Reconstruction via the Hartley Transform", *South African Computer Journal*, Pages 36-42

[Opp78] Oppenheim A.V. (1978), "Applications of Digital Signal Processing", Prentice-Hall, ISBN 0-13-039115-8

[Opp89] Openheim A.V. and Schafer R.W. (1989), "Discrete-Time Signal Processing", Prentice-Hall (New Kersey), ISBN 0-13-216292-X

[Par86] Parsons T. (1986), "Voice and Speech Processing", McGraw-Hill Inc., New York

[Pir84] Pirani G. and Zingarelli V. (1984), "An Analytical Formula for the Design of Quadrature Mirror Filters", *IEEE Transaction of Acoustics, Speech, and Signal Processing*, Vol. ASSP-32, No. 3

[Pre88] Press W.H., Flannery B.P., Teukolsky S.A., and Vetterling W.T. (1988), "Numerical Recipes in C", The Art of Scientific Computing, Cambridge University Press, ISBN 0-521-35465-X

[Ran87] Randall R.B. (1987), "Frequency Analysis", Brüel and Kjær, ISBN 87-87355-07-8

[Ros76] Rosenberg A.E. (1976), "Automatic speaker verification - a review", *Proceedings IEEE*, Vol. 64, No. 4, Pages 475-487

[Row92] Rowden C. (1992), "Speech Processing", McGraw-Hill Book Company, ISBN 0-07-707324-X, London

[Rus92] Ruskai M.B., Beylkin G., Coifman R., Daubechies I., Mallat S., Meyer Y. and Raphael L. (1992), "Wavelets and Their Applications", ISBN 0-86720-225-4

[Sch75] Schafer R.W., and Rabiner L.R. (1975), "Digital Representation of Speech Signals", *Proceedings IEEE*, Vol. 63, Pages 662-677, April 1975

[Scu92] Scully C., Castelli E., Brearley E. and Shirt M. (1992), "Analysis and simulation of a speaker's aerodynamic and acoustic patterns for fricatives", *Journal of Phonetics*, Vol. 20, Pages 39-51

[Sim90] Simoncelli E.P. and Adelson E.H. (1990), "Non-separable Extension of Quadrature Mirror Filters to Multiple Dimensions, *Proceedings of IEEE*, Vol. 78, No. 4

[Sle90] Slezak E., Bijaoui A. and Mars G. (1990), "Identification of structures from galaxy counts: use of the wavelet transform", *Astronomy and Astrophysics*, Vol. 227, Pages 301-316

[Smi86] Smith M.J.T. and Barnwell T.P. (1986), "Exact Reconstruction Techniques for Tree-Structured Subband Coders", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 3

[Sol81] Soli S.D. (1981), "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation", *The Journal of the Acoustical Society of America*, Vol. 70, No. 4

[Ste72] Stevens K.N. and Klatt D.H. (1972), "Current models of sound sources for speech", *Ventilatory and Phonatory Control Systems: An International Symposium*, edited by Wyke B.D. (London), Pages 279-291

[Str89] Strang G. (1989), "Wavelets and dilation equations: A brief introduction", *SIAM Review*, Vol. 31, No. 4, Pages 614-627

[Tay83] Fred J. Taylor, "Digital Filter Design Handbook", Marcel Dekker Inc. (New York), ISBN 0-8247-1357-5

[Tet91] Tetschner W. (1991), "Voice Processing", Artech House (Boston), ISBN 0-89006-468-7

[Vai85] Vaissière J. (1985), "Speech Recognition: A Tutorial", Computer Speech Processing, Fallside F. and Woods W.A. (ed), Prentice/Hall International, ISBN 0-13-163841-6

[Vet84] Vetterli M. (1984), "Multi-Dimensional Sub-Band Coding: Some Theory and Algorithms", *Signal Processing*, Vol. 6, Pages 97-112

[Wai88] Waibel A. (1988), "Prosody And Speech Recognition", London, Pitman, ISBN 0-934613-70-2

[Wat92] Waters G. (1992), "Speech Production and Perception", Speech Processing, Rowden C. (ed), McGraw-Hill International (UK), Maidenhead

[Wea83] Weaver H.J. (1983), "Applications of Discrete and Continuous Fourier Analysis", John Wiley and Sons Inc., ISBN 0-471-87115-X.

[Woo86] Woods J.W. and O'Neil S.D. (1986), "Subband coding of images", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-34, No. 5, Pages 1278-1288

Appendix A

Hardware and Software, and Sampling Processes

This chapter is dedicated to clarifying the conditions under which the research was accomplished and what hardware and software issues were considered. The purpose of Appendix A is to discuss briefly the hardware and software used in all stages of the research as well as mentioning the sampling strategies. Everything from the speech sampling hardware and software right through to the preparation of this document is detailed.

A.1 Computer Hardware used for Sampling and Processing

Most of the hardware used was provided by the Computer Science Department of Natal (Durban) and proved sufficient for most of the work. Some active filtering hardware was constructed by the author towards the beginning of the research. Some sampling hardware was never seen by the author as only the samples were obtained. This 'blind sampling' occurred while exploring inter-disciplinary uses of the Dominant Scale Transform, with the Physics Department.

Acoustic signals were recorded using an off-the-shelf IBM compatible 386 with a 16-bit SoundBlaster analogue-to-digital card, and a cheap 600Ω microphone with maximum frequency response of 9 kHz. The high-frequency cut-off could be viewed as a low-pass filter and therefore no active or passive anti-aliasing filters were included [Ham83, Lan88]. All signals were sampled at exactly 22.050 kHz. No attempt was made to filter or eliminate back-ground noise, although the room was kept quiet apart from ambient computer noises.

For research into the generating of music scores from an acoustic piece of music, the author used a music keyboard which stored pre-recorded musical instruments. We should emphasise that the sounds were not synthesised, but rather consisted of actual pre-recorded sounds of each instrument that had been digitally recorded. These sounds were played back at different amplitudes, pitch, and sustain to obtain the desired effect.

Sampling Hardware Specifications	
Computer Hardware	386DX-40 CPU, 8 Meg of RAM, VGA monitor, and 170 Meg hard-drive
ADC Hardware	SoundBlaster
ADC precision	8 bits
Sampling Rate	22.050 kHz
Microphone	Cheap Ammp 600Ω dynamic microphone
Microphone cut-off frequency	Approx. 9 kHz
Music Keyboard	Yamaha PSR-31 Pulse Code Modulation Keyboard

Table A-1: The hardware used during the sampling.

Our first attempt at sampling used an off-the-shelf IBM compatible PC-30 A/D card in an Intel 286 computer. The author built a front-end high quality pre-amplifier, attenuated by a low-pass active filter. This comprised of an active sixth-order slight-dips low-pass filter with approximately a 3.4 kHz cut-off frequency. A sixth-order active filter was used to improve the roll-off characteristics; -36dB/octave roll-off [Ham83]. The signals were sampled at 10 kHz. Unfortunately, the higher frequencies were found to be significantly reduced in amplitude. We therefore concluded that the low-pass cut-off frequency was too low and therefore the sampling rate was increased. All the samples appearing throughout this text have been sampled at 22.050 kHz using a microphone which has a significant roll-off around 9 kHz. We found it unnecessary to include either an active or passive low-pass anti-aliasing filter.

A.1.1 The Analogue-To-Digital Converter

The Analogue-to-Digital Converter (ADC) is the piece of hardware which links the analogue world outside the computer to the digital world inside. If we sample the digital output fast enough, and then use a Digital-to-Analogue Converter (DAC), we could accurately recreate the sound. It is this process which is used in compact disc (CD) players. The appropriate sampling rate for any given signal is discussed later in the appendix. We shall always assume linearity between the analogue and digital representations.

A.1.1.1 Digital Representation

The norm for the modern desktop computer has become a 32-bit system. Representing analogue values in 32-bits has the advantage that the resolution is extremely fine, in fact, steps would be practically undetectable. However, we commonly use 16-bit sampling as sufficient resolution exists and we achieve a halving of the amount of data to store and process. We have chosen to use 8-bit sampling. Throughout the research, we tried to work with a minimal system, for example, cheap microphones etc. Therefore our 8-bit (byte) digital representation has a range of [0..255] with 128 representing zero amplitude. A 16-bit (integer) has range [-32768..32767] with 0 representing zero amplitude. Although the 16-bit system effective doubles the amount of data to be processed, no speed penalty is incurred as the total number of samples is not dependent on the sampling resolution. The current implementation uses 8-bit sampling and 16-bit internal processing.

A.2 Computer Hardware used for Sampling and Processing, and Writing of the Manuscript

A.2.1 The Sampling Process

The recording of sound waves relies on pressure changes on a diaphragm, which induces an electrical current. It is essentially this current which is detected and recorded. These waveforms can be represented mathematically by functions of the form $f(t)$ where the variable t relates to time. The interpretation and storage of signals involves a discretisation of the original analogue signal. This is achieved through sampling and the conversion process is known as *analogue-to-digital conversion*, or simply A/D conversion.

The discretisation process obviously results in a (slight) degradation of the original signal. A far more serious potential for loss of information of the incoming signal is that of *aliasing*. Assuming a constant *sampling period* Δt between each sample taken, the *sampling frequency* f_s would be given as $f_s = 1/\Delta t$. The minimum sampling frequency of a signal before aliasing occurs is known as the *Nyquist Frequency* and is exactly $f_s/2$. The 'highest significant frequency component' in the sampled signal $f(t)$ should therefore not exceed the Nyquist frequency. A component of the incoming signal having higher frequency than the Nyquist frequency will result in an under-sampling of the incoming signal; the outcome being aliasing.

We now state the *Dirichlet conditions* which, if satisfied, guarantee that a function f will have a Fourier series frequency content [Wea83]. The Dirichlet conditions are:

1. f is periodic with period T; that is, $f(t+T)=f(t)$.

2. f is bounded.

3. In any one period, the function may have at most a finite number of discontinuities and a finite number of maxima and minima.

For the duration of the research presented in this thesis, only regularly sampled signals were used, that is, signals in which the Δt between adjacent samples stays constant throughout the entire signal and which satisfy the Dirichlet conditions. It was not necessary to normalise the input signals, although this may be required if the system is to be implemented commercially.

The entire sampling process was performed on an MS-DOS Version 5.00 operating system, with the Microsoft Windows 3.1 Graphical User Interface running the MS-WaveEdit Version 1.0c sampling software. The platform performed surprisingly well at a sampling rate of 22.050 kHz.

A.2.2 Graphing Tool Details

All plots were created with a ShareWare program called FastGraph. Unfortunately, it was only in the latter stages of the write-up that it became apparent that very large data series could not be handled by FastGraph. This is a direct result of segments used in MS-DOS implemented on Intel CPUs. The difficulty was solved in two ways, albeit neither of them optimal:

- All the speech signals have had every other sample removed to halve the size of the data series. The axes labelling has not changed as a result; however, some of the definition of the high frequency regions may be lost. This was not found to be of any significance.

- A few speech signals, although already reduced in size by the action above, were still too large to handle and therefore a small amount of truncation at their endpoints was effected.

Many of the graphs in the text have their axes normalised. For example, in the processing of speech signals, the rate at which the word was spoken should not affect the recognition system adversely. We therefore normalise the axes on the graphs.

A.2.3 Miscellaneous Software

The research (except for the sampling) was performed on an IBM OS/2 2.1 operating system platform having a 386DX-40 CPU, 8 Meg of RAM, VGA monitor, and a 170 Meg hard-drive. The OS/2 operating system allowed for pre-emptive multi-tasking which permitted the editing of documents while results were being generated in a background process. The entire Ph.D. was written in Microsoft's Word for Windows 2.0a. The Microsoft Word's English (UK) spelling checker was used throughout the document and, to the best of our knowledge, the document is free from spelling errors.

All Fourier Transforms and interpolations were calculated using Mathematica 2.0 for MS-DOS 386/7 which is Copyright 1988-91 Wolfram Research Inc.

Sampling Software Specifications	
Sampling Operating System	MS-DOS Version 5.00 and MS-Windows 3.1
Sampling Software	MS-WaveEdit Version 1.0c

Table A-2: The software used during the sampling phase.

Appendix B

Music Tables

Note	Hz	Note	Hz
G	392.00	C$^{\#}$	554.37
G$^{\#}$	415.30	D	587.33
A	440.00	D$^{\#}$	622.25
A$^{\#}$	466.16	E	659.26
B	493.88	F	698.46
C	523.25	F$^{\#}$	739.99

Table B-1: One octave of an equal tempered chromatic scale.

Appendix C

Glossary

AC:	Alternating Current
ADC:	Analogue-to-Digital Converter
AKC:	Angela Kay Cooper
CD:	Compact Disc
DAC:	Digital-to-Analogue Converter
DC:	Direct Current
DFT:	Discrete Fourier Transform
DIC:	David Ian Carson
DSA:	Dominant Scale Algorithm
DST:	Dominant Scale Transform
FFT:	Fast Fourier Transform
FHT:	Fast Hartley Transform
FT:	Fourier Transform
FWT:	Fast Wavelet Transform
HG:	Hilton Goldstein
IDFT:	Inverse Discrete Fourier Transform
OS/2:	Operating System/2
PC:	Personal Computer
SDST:	Spectrogram Dominant Scale Transform
STD:	Standard Deviation
STFT:	Short-Time Fourier Transform
WT:	Wavelet Transform

Theorem: Given any real, non-italic Times New Roman *x* and any non-zero, non-italic Times New Roman *n*, we have:

$$\frac{\mathrm{sin\ x}}{\mathrm{n}} = 6$$

Proof: [Hil&Viv et al 93]

$$\frac{\mathrm{si}\cancel{\mathrm{n}}\ \mathrm{x}}{\cancel{\mathrm{n}}} = 6 \qquad\qquad \square$$

“Man achieves by accomplishing that which is challenging, not that which is simple.” - *Hilly '94*