The Multiresolution Fourier Transform

and its Application to

Polyphonic Audio Analysis

Edward R. S. Pearson B.Sc.

Department of Computer Science

A thesis submitted to
The University of Warwick
for the degree of
Doctor of Philosophy

September 1991

Contents

List of Figures

Acknowledgements

This work was supported by UK SERC and Solid State Logic Ltd. It was conducted within the Image and Signal Processing Group in the Department of Computer Science at The University of Warwick, UK.

Many people have contributed to this work, in particular, the staff and students of the department. Special thanks goes to Jeff Smith, Rod Moore and Roger Packwood who gave invaluable software support and assistance with hardware construction.

Great thanks must go to my friends and colleagues for their stimulation and support — Abhir Bhalerao, Andrew Calway, Simon Clippingdale, Roddy McColl, Andrew Davies, Hayley Ryder, Martin Todd, Philip Underdown, June Wong, the members of Warwick University Caving Club and many others.

I am particularly indebted to my supervisor, Dr. Roland Wilson. Without his great knowledge, understanding and unending stream of ideas, this work would not have been possible.

Finally I would like to thank my mother for her endless love, support and patience, Catherine for giving me the inspiration to carry on when times were hard and Cho for making a dream come true.

Chapter 1

Introduction

1.1 Artificial Hearing

Many people listen to, or at least hear, some form of music almost every day of their lives. However, only some of the processes involved in creating the sensations and emotions evoked by the music are understood in any detail. The problem of unravelling these processes has been much less thoroughly investigated than the comparable topics of speech and image recognition; this has almost certainly been caused by the existence of a greater number of applications awaiting this knowledge. Nevertheless, the area of music perception has attracted some attention over the last few decades and there is an increasing interest in the subject largely arising from the availability of suitably powerful technology. It is becoming feasible to use such technology to construct artificial hearing devices which attempt to reproduce the functionality of the human auditory system. The construction of such devices is both a powerful method of verifying operational theories of the human auditory system and may ultimately provide a means of analysing music in more detail than man. In addition to the analytical benefits, techniques developed in this manner are readily applicable to the creative aspects of music, such as the composition of new music and musical sounds.

1.2 The human auditory system

The human auditory system extends from the outer ear (pinna) through a complex and highly sensitive set of mechanical and chemo-electrical systems to the temporal lobe of the brain. The physical layout is normally divided into the outer, middle and inner ears; the auditory nerve and central auditory pathways. The ear itself, a highly complex system comprising many delicate components, not only serves to convert sound waves in the atmosphere to nerve impulses but also includes the sensory organs of balance. Though the processing performed by the ear on the acoustic signal has been the subject of a long, and sometimes heated, debate, it is now generally accepted that it operates as both a temporal and frequency analyser. The inner ear performs the conversion from mechanical vibrations to nerve impulses by means of the cochlea, a spiral tube divided in two along its length by the basilar membrane. The tube becomes smaller in diameter along its length so that, as sound waves pass along it, different frequencies resonate in different regions and excite tiny hair cells on the membrane. At low frequencies at least, these hair cells can respond fast enough to respond synchronously with the extreme excursions of the signal. This topic is the best documented area of hearing, having a large body of work associated with it, see [Gel81] or [Nor70] for thorough surveys. After the cochlea, the signal passes into the auditory nerve and our degree of knowledge of the physical aspects of its subsequent processing and interpretation starts to decrease fairly rapidly. Experimentation by many workers has, however, suggested the qualities of the signal which are measured and the types of features recognised by the subsequent processing stages.

Figure 1.1 represents the human auditory system as an information processing hierarchy. At the lowest level the musical signal is in its most fundamental form, an acoustic wave. As the signal passes up the hierarchy it is refined and interpreted making its form and the quality of information it carries increasingly abstract. The most clearly defined percept for the listener is the symbolic form of the signal. The listener perceives a stream of events within the music [MB79], which predominantly correspond to the sounds produced by each of the performing instruments. Of these

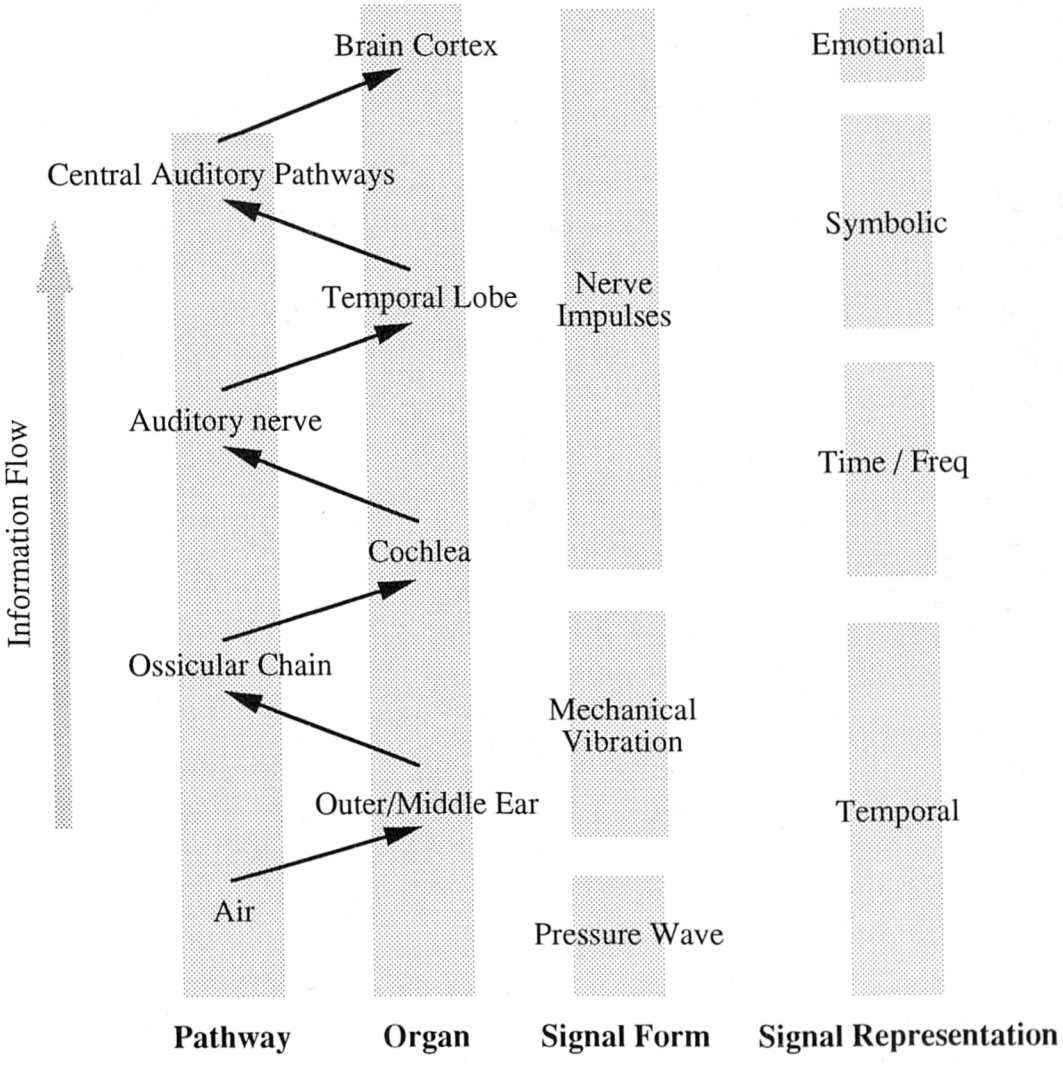

Figure 1.1: The Human Auditory System as an Information Hierarchy

'events', some are perceived as having a distinct pitch and are commonly referred to as notes. The number of properties which may be attributed to notes, as demonstrated by the richness of language used to describe them, is vast. For the purposes of analysis, however, only four are typically discussed, namely: pitch, loudness, duration and timbre. In the case where a single note has been isolated, pitch and loudness are primarily functions of the frequency and amplitude of the signal, respectively. Duration is fairly simply related to the physical duration of the signal, though exactly which facets of the signal correspond to timbre is largely unknown. Part of this problem is that the term is fairly loosely defined; W. Dixon Ward in [Nor70] says that the term has become a 'wastebasket' category, "if two sounds are 'different' though having the same pitch and loudness, then they must differ in timbre". Classical (now obsolete) theory speculates that this quality is related to the spectral distribution of the note; while certainly contributing, the simple experiment of observing the great difference in timbre of playing a sound backwards, via a tape machine, suggests that temporal aspects of the signal also play an important role. J. M. Grey in [Gre75] devised a timbral space in which the distance between two points corresponded to the perceived difference in timbre between two notes. However, the space was produced by a multi-dimensional scaling algorithm run on the results of a set of listening tests and so, unfortunately, the relationships between the axes of the space and any physical attributes of the test signals could not be deduced.

The study of the human perception of pitch is possibly the most thoroughly investigated areas of psychoacoustics. For simple (monophonic) signals, pitch is the psychological correlate of the frequency of the signal, but the relationship is far from simple. The method by which the auditory system determines the pitch of a stimulus is one of the oldest and most heated areas of debate in psychoacoustics, centering around a percept known variously as fundamental pitch, residue pitch, virtual pitch, or periodicity pitch. The percept is the *single* pitch identified from a tone containing one or more frequency components (a tone complex). The exact pitch perceived depends on many factors including the absolute and relative frequencies of the tone's partials (see below), their relative

amplitudes and on the listener themselves. One of the more important findings of this investigation is that the tone may not have a component at the frequency of the pitch it is perceived to have; hence the term *virtual* pitch. Many thorough surveys are available on this subject including [Sma70] and [Gel81], while a more musical look at the topic can be found in [War70].

Notes can be broken down into a set of partials, each a pseudo-sinusoidal function having a single frequency. For a note to have a clear pitch of N Hz. the partials with the largest magnitudes have frequencies of N Hz, $2N$ Hz, $3N$Hz. etc. These partials are normally referred to as the harmonics of the note, the others sometimes being referred to as the inharmonic partials. As the relative magnitude of the harmonics decreases with respect to the inharmonic-partials the pitch of the tone becomes decreasingly distinct. The tones produced by bells have inharmonic-partials with significant magnitude. For the notes analysed in this work, it is assumed that the magnitude of the inharmonic-partials is so small that they can be ignored. For this reason, the terms harmonic and partial are used interchangeably.

1.3 Objectives of this work

The work described in this thesis leads to a system which can interpret a limited set of signals and produce output at the level of symbols which correspond to the notes of a score. The term *automatic transcription system* [1] has become applied to such such systems, since the symbolic representation produced is similar to the information represented on a musical score and the process can be considered equivalent to writing down the score on hearing a piece of music.

The main objectives of this work are:

1. Investigate some aspects of the nature of musical signals in the context of signal processing and pattern recognition.

[1] where *transcription* is used in the literary rather than the musical sense.

2. Develop a signal representation suited to the analysis of musical signals.

3. Devise analytical techniques which make the best use of the devised representation.

4. Build an example application using the representation and the techniques in order to assess their potential.

1.4 Automated Transcription

In order to move towards a methodology for building an automatic transcription system, the problem is first discussed in general terms.

Figure 1.2 shows the magnitude of a short segment of a musical signal with respect to time. Figure 1.3, which is the musical score which would have been performed in order to produce the signal.

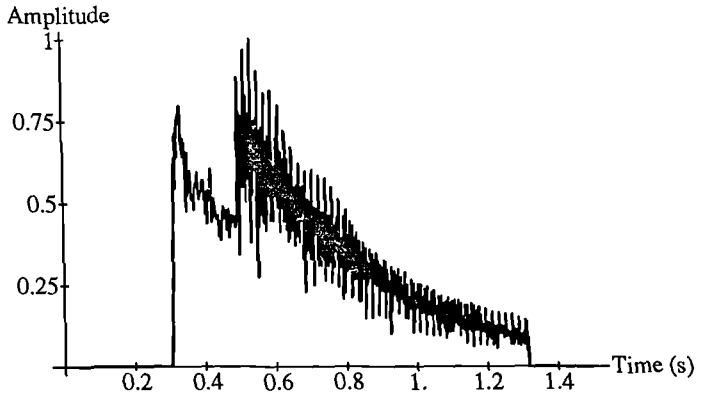

Figure 1.2: Magnitude of Two Piano Note Audio Signal

The score is a symbolic description of the music in the sense that it is constructed from a set of discrete elements (the symbols) which are in this case, of course, its notes. The score too uses its

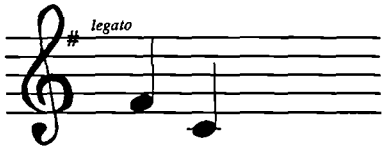

Figure 1.3: Score for Two Piano Note Signal

horizontal scale to measure time but represents frequency or pitch on its vertical scale; which is a far more perceptually important parameter of the music than its magnitude with time. In addition, the durations of the notes are represented by the form of the symbols used.

A note can be thought of as a special kind of sound, though it is not possible to define the precise point at which a sound becomes a note. Both are experienced as a single percept, but that described as a note, by a typical listener, will have a definite pitch associated with it as well as a clear starting time and, often less distinctly, a finishing time. A listener could judge which of two sounds started first but could only attempt to place them above or below each other in pitch if they were notes.

A simple working definition of the task to be implemented by a music transcription system can now be proposed:

Definition 1.1 *The requirement of an automatic music transcription system is to produce a 'score-like' representation of a piece of music given only a description of the amplitude of its acoustic signal with respect to time.*

The minimum information required in the output data consists of estimates of the onset time, frequency and duration of each of the notes in the signal.

The techniques developed in this work are specifically aimed at polyphonic signals. A polyphonic, rather than monophonic, signal is one where more than one note occurs simultaneously. Typically the classification is based on the number of instruments playing and how many notes they can sound at any time. The oboe, for instance, has only one sound producing element, its reed, and so is classed as monophonic. An instrument such as a piano has a set of strings for each note and is

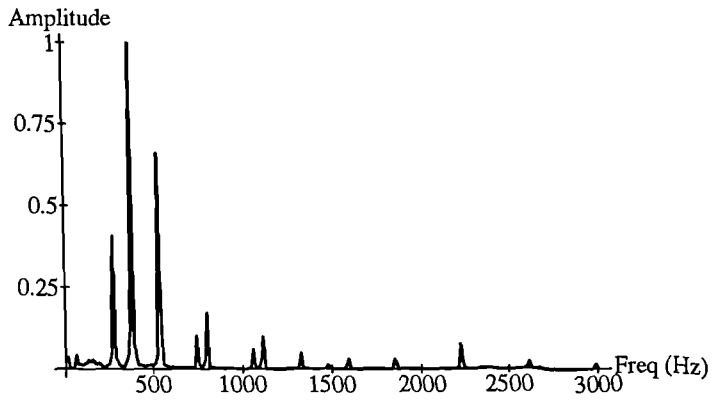

Figure 1.4: Magnitude of Fourier Transform of Two Piano Note Signal

not restricted to the number of notes it can play at once; it is polyphonic. From a signal processing point of view, the situation is not quite as clear. The simple two note signal presented above would probably be classified as monophonic since only one note is 'played' at any time, however the first note is still decaying when the second starts (legato), and so must be considered polyphonic for the purposes of signal analysis. Much previous work has been applied to monophonic sources. While not without value, the techniques developed are often dependent on this monophonicity and so can never be applied in polyphonic situations. The work presented here has tried to avoid such limitations.

The structure of the music is clearly represented by its score in Figure 1.3: there are two notes at different frequencies. The signal, on the other hand, is a continuous function and so has a far less transparent structure. Having said that, in such a simple case, it is easy to postulate that each of the peaks in the signal correspond to the onsets of the notes. For this example, it would be a fairly trivial exercise to estimate manually the onset time of each note with no more than a ruler. However, it would be extremely hard, if not impossible, to determine the pitch of both notes. The problem is

twofold: the signal time plot has no explicit frequency scale and so this parameter would have to be determined implicitly from the temporal behaviour of the signal and after the onset of the second note, the signal is the sum of two component signals which would have to be separated from each other before such analysis could take place. Such separation of a signal into a set of component signals is know as *signal segmentation*; to perform this on a given class of signals requires the definition of a *segmentation strategy* which has been discussed in general terms by several authors including Harlick [Har83] Wilson and Spann [WS87b]. The deduction of a such a strategy for music transcription forms the basis of Chapter 4.

The problem of not having an explicit frequency scale for the raw data can be approached by applying some transform to the data which results in an alternative representation of it. This is discussed at length in the next chapter; for now observe Figure 1.4, which is the magnitude of the Fourier Transform (FT) of the signal. This representation has axes displaying frequency and amplitude. Clear peaks can be seen corresponding to the harmonics or partials of the notes. The frequencies of the notes could be obtained from these peaks but notice that there exists the problem that the harmonics are intermingled and need to be assigned to notes which is, again, a signal segmentation problem. However, the more fundamental problem with this form of the signal is that there is now no time axis; it is therefore no more use for deducing all of the parameters of the notes than Figure 1.2, since it is impossible to observe either the onset times or durations of the notes.

The solution to this dilemma can, at least partly, be resolved by the use of some signal representation which has both time *and* frequency axes. An example of such a representation is shown in Figure 1.5, which is a Short Time Fourier Transform (STFT). This is a three dimensional representation, with axes of time, running front to back, frequency left to right, while the amplitude at each point is represented by the height of the surface. The shading in this and other 3D plots used in this thesis serves to reinforce the perspective of the surface rather than indicate any value. The signal represented is, again, the two piano notes. The plot consists of a small number of ridges

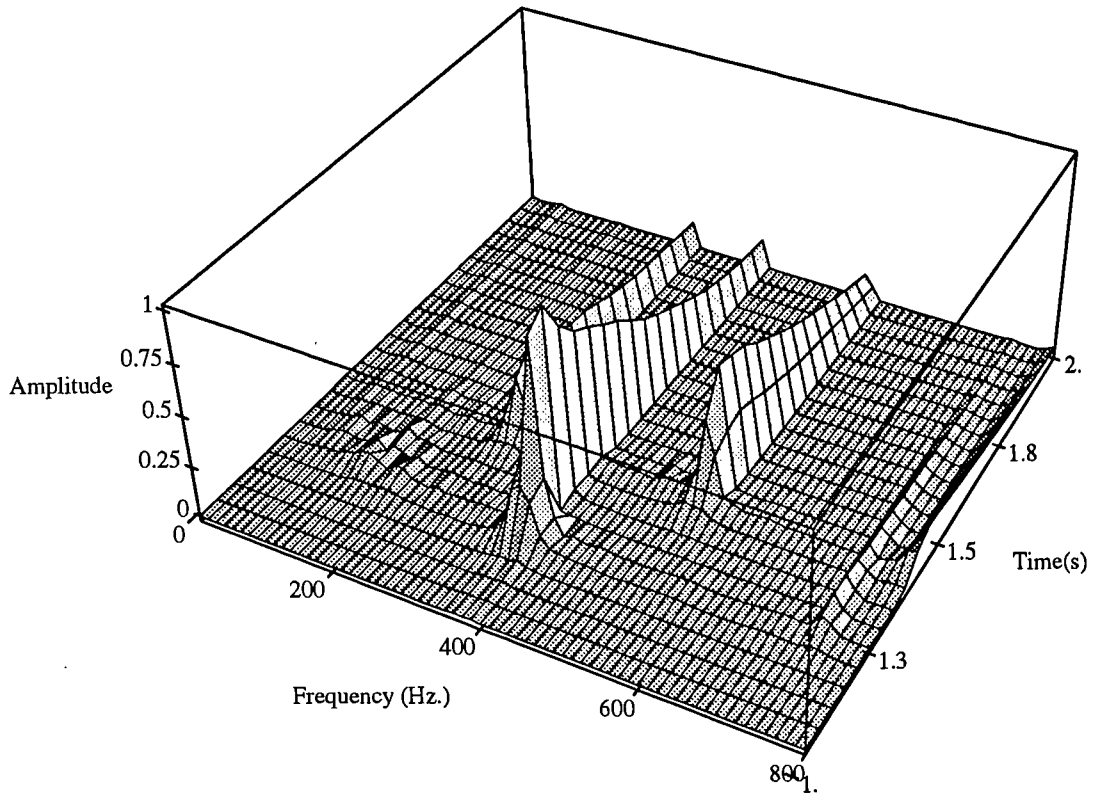

Figure 1.5: A combined time-frequency representation of two piano notes

running front to back which correspond to the partials (see below) of each note. The front edges of the ridges lie at the onset times of the partials and closer inspection reveals that there are two sets of ridges grouped by onset time. Each note corresponds to one of these sets. The first set of ridges (toward the front) is more widely spaced than the second, corresponding to the pitch of the first note being higher than the second. The basic parameters of the music can now be observed directly from the signal representation. There are clearly two sets of partials, they occur at different times and the first is higher in frequency than the second. The argument here, all be it fairly heuristic, is simply that this kind of time-frequency representation is clearly more suited to the analysis of musical signals than a representation describing it exclusively in terms of time or frequency. For these reasons, it has become a standard technique to use time-frequency representations (with the Short-time Fourier Transform being the most common) for the analysis of audio and many other signals. Various forms of time-frequency representations and their relative merits are discussed in Chapter 2.

1.5 Previous Work

The most formative work in the area of polyphonic music transcription is undoubtedly that of J.A.Moorer, both in his thesis [Moo75] and later works e.g. [GM77] and [Moo78]. The thesis, "On the Segmentation and Analysis of Continuous Musical Sound by Digital Computer", identifies many of the problems associated with the task and includes a review of various signal processing techniques which were available at the time. The system presented used a directed bank of hetrodyne filters to track partials. The algorithm required that the key signature of the piece be determined at an early stage in order to direct the filterbank. The source material was limited to two note polyphony, the lines played were not allowed to cross in frequency and the interval between the lines had to be between a minor third and minor seventh. The system included techniques for presenting the results in the form of a musical score and attempted to closely match the original score.

Work by Charles Watson in [Wat86], using a variety of heuristic techniques, obtained accurate results from synthesised and natural signals. The algorithms developed need to be tailored interactively to suit the characteristics of the particular signal being analysed, it is not clear whether they could be generalised to cope automatically with a range of input signals.

A variety of techniques have been described by Chafe et al in [CMR82, CJK+85] and [SS86] from the Centre for Computer Research in Music and Acoustics (CCRMA). These techniques have been used as tools forming part of an interactive signal editor, but it is not known if they have, to date, been successfully integrated into a fully-automated analysis system.

Of particular interest is a recent work by Serra [Ser89], though being oriented towards sound manipulation and resynthesis rather than transcription, it does demonstrate the power gained by using a feature based analysis method. Very briefly, the system generated a pseudosinusoidal expansion of the signal using a short-time Fourier transform. The representation could then be used to drive an additive synthesiser, the output of which could be combined with the residual of the original to produce realistic reconstructions. The system is oriented towards monophonic sound sources but can cope with the kind of 'overlap' polyphony described above. The user is required to know some of the parameters of the music as it is necessary to enter these before the analysis is performed. There is also an interactive stage after the automatic analysis when the accuracy of the partial tracks detected can be improved by the user.

Work on the synthesis of realistic sounds has much to contribute to the successful analysis of natural signals. Only by accurate modelling of the sound generating processes of natural instruments can an understanding be gained of the signals they produce. An ideal analysis system should be able to recognise the individual subtleties of the wide range of sound sources in common use today while retaining the ability to identify their common features. In recent years sufficient computational power has become available to allow highly complex models of the physical elements of natural instruments to be used for synthesis rather than the simple oscillator and frequency modulation (FM)

techniques used by currently available commercial synthesisers. Such methods have been applied to the modelling of reed and string instruments [AR85, ACR87, Smi87, ACE88] and the human vocal tract [RDP87], most notably as part of the CHANT project [Rod84, RPB84]. The models used in this thesis are simple by comparison though there is no fundamental reason why more complex models could not be used within the analysis framework presented. The limitations of the signal model used are discussed in the final Chapter.

1.6 Mathematical Notation

The mathematical notation used in this thesis makes use of a great many superscripts and subscripts. In order to aid the reader, an attempt has been made to use this notation in a consistent manner. A particular letter is associated with each value to be represented e.g. t for a time. The letter is emboldened when it represents a vector e.g. $\mathbf{t}$ might be a vector of times. Elements of such vectors are referenced using a subscript e.g. for a single dimensional vector, t_i would be the ith element of $\mathbf{t}$. Elements of multidimensional vectors are referenced by means of a comma separated list of subscripts.

Sometimes there are occasions when letters are required to represent several related values. In these situations a tag letter is attached as a subscript, or if a subscript already exists (e.g. a vector dereference) as a superscript. For instance f_p might be a frequency associated with a partial while f_n might be a frequency associated with a note which may also have an associated time t_n. Superscripts are also used in their normal role of exponentiation and it is hoped that the context will make the intended meaning clear. This use of superscripts and subscripts can be applied recursively so that each may have superscripts and subscripts of their own, though excessive levels of recursion have been avoided.

1.7 Equipment Used

The experimentation and software development for this work was performed on several Sun Microsystems computers. Analogue to digital and digital to analogue conversion of the audio signals was via equipment kindly supplied by Solid State Logic Limited of Oxford, U.K. who also partly funded the work. A sampling precision of 16 bits and a sampling rate of 48kHz was used for all signals. Test signals were synthesised using the CSound package from MIT, while natural sound sources were provided by the McGill University Master Samples Compact Discs and a variety of other prerecorded sources.

1.8 Thesis Organisation

The following Chapter discusses the properties of various combined time-frequency representations leading into a description of the Multiresolution Fourier Transform (MFT) which forms the basis of the rest of the work. Chapter 3 discusses the more practical aspects of the MFT including its implementation and application to musical signals. This discussion is proceeded by the presentation of some transforms of simple signals to allow familiarisation with the qualitative aspects of this signal representation. Chapter 4 then turns to the modelling of harmonic based musical signals within the MFT which is developed into a set of feature detection algorithms in chapters 5 and 6. The results of applying these algorithms to various pieces of music are presented in Chapter 7, while conclusions, suggested improvements and further work are covered in Chapter 8.

Chapter 2

Representations of Audio Signals

2.1 Introduction

This chapter comprises two main parts. After a brief discussion, there is a review of several signal representations which have been applied to the analysis of audio signals: the next section then introduces and defines the *multiresolution Fourier transform* (MFT) which forms the basis of the work presented in the subsequent chapters.

It is important to distinguish between a signal and a representation of that signal. The representation can be considered to be a view of the signal by some observer. A given signal may have many representations: many different views of the same data. For one dimensional signals, such as audio, the most common representation is that given by observing the signal in the time domain — other representations can be obtained by applying some signal transform. Different representations can give highly dissimilar views: some aspects of the signal may be revealed in one representation and hidden in another, as was demonstrated in Chapter 1.

Any system which processes audio signals must represent those signals internally in one or more ways. Simple systems which only perform modifications on an audio signal (e.g. an equaliser), may only use only the time domain description: the modifications they make are easily implemented

in the time domain. An audio analysis system, which produces a symbolic output, rather than a modified signal, typically transforms the signal from the time domain into some more suitable representation, on which it then operates. The suitability of the representation depends upon the goals of the system. Where those goals are to produce a *perceptually relevant* set of symbols from the input signal, then it can be seen that the algorithms that the system must implement will be simplified by allowing them to operate on a signal representation which describes the signal in similar terms to those in which it is perceived. This similarity is of particular importance in an *interactive* analysis-synthesis system [CMR82, Ser89] where the transformed signal may be directly observed and possibly modified by the user. As was discussed in the previous chapter, it is generally agreed that the best representation for this type of work is one which incorporates aspects of both time and frequency and these are commonly referred to as *combined* or *conjoint* [Dau88b] representations.

2.2 Some commonly used representations

A wide variety of time-frequency representations have been investigated by many different workers and, as a result, there has been much literature published on the subject. Several comprehensive surveys of the topic have been published; a most exhaustive review has recently been given by [Coh89] and there is little point going into as much detail here. All combined representations fall into one of a few broad categories and examples from each of these are discussed in the following sections, with an emphasis on those which have been applied to musical signals.

2.2.1 Linear Transforms

A transform G is termed *linear* if it has the following property. If some signal $v(t)$ can be described as a linear combination of signals,

$$v(t) = \sum_{i=1}^{N} k_i v_i(t) \tag{2.1}$$

where k_i are constants, then its transform $G(v)$ satisfies

$$G(v(t)) = \sum_{i=1}^{N} k_i G(v_i(t)) \tag{2.2}$$

The transform of the composite signal is the sum of the appropriately weighted transforms of the component signals. Equation 2.1 is a suitable model for musical signals, which, typically, consist of individual sounds 'mixed' together. A linear transform ensures that this form is retained in the transformed domain, clearly simplifying any analysis system based on it.

The Fourier Transform

The single most important transform in signal processing is the *Fourier transform* (FT) which relates the time and frequency domains. The Fourier transform $V(\omega)$ of a signal $v(t)$ is defined by

$$V(\omega) = \int_{-\infty}^{\infty} v(t) e^{-j\omega t} dt \tag{2.3}$$

The inverse transform is then defined as

$$V(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} V(\omega) e^{j\omega t} d\omega \tag{2.4}$$

A similar transform can be defined for discrete signals and is referred to as the *discrete Fourier transform* (DFT).

This transform is of little use for any practical analysis system, as the infinite integration involved means that it must be evaluated for all time. Rather, the transform is a mathematical ideal [Sle76] most useful as a base on which to build other transforms. For practical purposes the data are normally multiplied by some window function, which is non-zero only within some limited range. This form is discussed in the next section.

The Short Time Fourier Transform

The short time Fourier transform (STFT) is often called the 'phase vocoder' in speech [Por76] and music applications [Moo78, Can80]. The name arises from the fact that the transform coefficients are complex; previous vocoders had not included phase information.

The forward transform is defined as [Por80]

$$V_2(t,\omega) = \int_{-\infty}^{\infty} h(t-\tau)v(\tau)e^{-j\omega\tau}d\tau \tag{2.5}$$

The function $h(t)$ is the *analysis window* and is chosen to be concentrated in both time *and* frequency. It is this *localisation* which means that $V_2(t,\omega)$ can be considered as a representation intermediate between $v(t)$ and $V(\omega)$. It can be seen from the STFT definition that it is a linear transform.

An inverse STFT can be defined. The original signal can be recovered using [Por80]

$$v(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(t-\tau)V_2(\tau,\omega)e^{j\omega t}d\tau dt \tag{2.6}$$

The function $f(t)$ is the *synthesis window* which, to ensure invertibility, must be related to the analysis window $h(t)$ by

$$\int_{-\infty}^{\infty} h(t)f(-t)dt = 1 \tag{2.7}$$

A discrete form of the STFT can be defined [Por80]. Given a time sequence, $x(n)$, then

$$X_2(n,m) = \sum_{i=-\infty}^{\infty} h(nR-i)x(i)e^{-j\frac{2\pi}{M}mi} \tag{2.8}$$
$$0 \le m < M$$

The indices n and m select the transform coefficients from a two-dimensional integral lattice covering the time-frequency plane. R and $2\pi/M$ are the sampling intervals in time and frequency respectively.

As for the continuous case, with appropriate choices of sampling intervals and windows, this form of the transform is invertible [Por80].

$$x(n) = \frac{1}{M}\sum_{k=0}^{M-1}\sum_{m=-\infty}^{\infty} f(n-m)X_2(m,k)e^{j\frac{2\pi}{M}km} \tag{2.9}$$

The windows must now satisfy an additional condition arising from the sampling process [Por80],

$$\sum_{s=-\infty}^{\infty} f(n-sR)h(sR-n+pM) = \delta(p) \quad \text{for all } n \tag{2.10}$$

Note that the representation has no redundancy when $R = M$.

The coefficients of the discrete STFT are best interpreted by considering rows and columns of the representation individually. A single column, $X_2(n_0, m)$, can, from Equation 2.9, be seen to be the DFT of the time windowed signal $h(n_0R - n)x(n)$. A single row is then a time sequence of coefficients, one from each column, with $m = m_0$.

$$X_2(n, m_0) = \sum_{i=-\infty}^{\infty} h(nR-i)x(i)e^{-j\frac{2\pi}{M}m_0 i} \tag{2.11}$$

which can be considered in terms of a convolution,

$$X_2(n, m_0) = h(nR) * (x(nR)e^{-j\frac{2\pi}{M}m_0 nR}) \tag{2.12}$$

This may be interpreted as the output of an *analysis filter* $h(n)$ operating on a frequency shifted version of $x(n)$ [Por80].

The summation limits in Equation 2.9 are given as $\pm\infty$, but in practice they depend on the analysis window, which is typically zero-valued outside some finite interval. This leads to the idea that efficient implementations of the STFT are possible using the fast Fourier transform (FFT) [Por76] and this is the technique normally used in audio analysis [Moo78]. More recently, multirate filter bank implementations [SB87] have been used, which permit the application of quadrature mirror filter (QMF) techniques [ECG76]. Such a scheme removes aliasing in the analysis stage allowing the original signal to be perfectly reconstructed simply and efficiently.

Clearly, an important consideration when implementing the STFT is the choice of window function. Many windows have been tried, including rectangular, Hamming, Hanning and Kaiser. These, and others, have been reviewed in many places e.g. [RG75] and [Har78] so their relative

merits are not considered here. Note, however, that the Kaiser window [F.74] is commonly used, e.g. [Ser89].

The Gabor Representation

Gabor (1947) did not agree with the prevailing idea that hearing was well represented by frequency based Fourier analysis, though the representation he proposed [Gab46] has more recently been shown to be related to the STFT [Bas81]. Gabor stated "...it is our most elementary experience that sound has a time pattern as well as a frequency pattern...". The Gabor representation is defined as an expansion of the signal

$$v(t) = \sum_{k,l=-\infty}^{\infty} C_{kl} g_{kl}(t) \tag{2.13}$$

where C_{kl} are the expansion coefficients and the basis functions, $g_{kl}(t)$ are time-frequency shifted versions of a Gaussian window $g(t) = e^{-\alpha t^2}$.

$$g_{kl}(t) = g(t - kT)e^{j(l\Omega t + \phi)} \tag{2.14}$$

$$T\Omega = \sqrt{\pi \alpha^{-1}} \tag{2.15}$$

where T, Ω, α and ϕ are all constants. Given the time and frequency dispersions of the basis functions and the sampling intervals T and Ω, each transform coefficient C_{kl} represents a region (Gabor called it a 'logon') of the time-frequency plane of size $T \times F$, at time kT and frequency lF.

A drawback of the representation is that the basis functions $g_{kl}(t)$ are not orthogonal and so calculation of the expansion coefficients, C_{kl}, is not straightforward, requiring either vast computational resources or some iterative approximation. Recent work has lead to more efficient techniques [Bas81] but these are still computationally expensive compared to FFT based implementations of the STFT. The Gabor Transform has found little application in computer music, apart from being the motivation behind the development of granular synthesis [Roa85], though more interest has been shown by the image processing community, e.g. [Dau88b].

Within the context of this thesis, the importance of Gabor's work was to identify that there is a fundamental limit to the resolution at which a signal may be simultaneously expressed in time and frequency. This arises from Heisenberg's uncertainty principle [Pap77] which had previously been applied to quantum mechanics, but has become a fundamental result in signal processing. The principle implies that the time t and frequency f of some phenomenon cannot be measured simultaneously with arbitrary accuracy. The phenomenon must be considered to be somewhat dispersed in both domains and so degrees of uncertainty must be associated with those measurements. Specifically, if Δt and Δf are the uncertainties in time and frequency respectively, then they are bound by

$$\Delta t \Delta f \geq 1 \tag{2.16}$$

This relation is reflected in the limitation in Equation 2.15 on the relative concentrations in time and frequency of the transform basis functions. The uncertainty restriction is not, of course, specific to the Gabor transform — it applies just as much to the window functions used for the STFT.

2.2.2 Non-Linear Transforms

Various attempts have been made to overcome the limitations imposed on the linear transforms by the uncertainty principle. These representations forsake the linearity (eqn. 2.2) of the methods described in the previous section in order to increase time-frequency resolution.

The most widely used of these transforms is the Wigner distribution (WD) [CM80a, CM80b, CM80c]. It is defined for a continuous signal $v(t)$ by

$$W_v(t,\omega) = \int_{-\infty}^{\infty} v\left(t+\frac{\tau}{2}\right) v^*\left(t-\frac{\tau}{2}\right) e^{-j\omega t} d\tau \tag{2.17}$$

The correlation involved in the calculation of the WD means that it is *bilinear*; the WD of a linear signal with just two components, $v(t) = v_1(t) + v_2(t)$, is given by

$$W_v(t,\omega) = W_1(t,\omega) + W_2(t,\omega) + 2\Re\left[W_{1,2}(t,\omega)\right] \tag{2.18}$$

where $\Re\left[W_{1,2}(t,\omega)\right]$ is the real part of the so called *cross* Wigner distribution of the signals v_1 and v_2.

As discussed above, linearity of a signal representation is an important consideration for its applicability to polyphonic musical signals; the bilinearity of the WD is a potential drawback in this application.

The WD suffers from a number of other disadvantages. These are, most notably

1. The WD cannot readily be inverted [CM80a]. Thus it is not suitable as the basis of analysis-synthesis systems which are typically required by computer musicians.

2. The definition of the WD must be somewhat compromised for discrete signals [CM80b]; this has lead to several different interpretations, most of which suffer from a certain amount of aliasing [CM83].

3. The infinite integration in equation (2.17) means that, for a practical implementation of the WD, some window function must be introduced [CM80b]. These windows suffer from the uncertainty problems described above, and so discrete forms of the WD provide little or no increase in resolution compared with their linear counterparts.

These disadvantages have restricted the use of the WD, although it has been applied in some areas of audio analysis [JK83, VKDV88].

2.2.3 Multiscale Techniques

Gabor's work in relating the uncertainty principle to signal processing, as described above, revealed the fundamental limit on the resolution which can be obtained by a windowed transform. Recent attempts to overcome this limitation have concentrated on multiscale approaches and these are often referred to as *wavelet transforms*, after [GM84]. Basically the idea is to represent the signal as an expansion of a *set* of functions (wavelets), rather than just one. These families of functions are

based on translations and dilations of some function, $g(t)$,

$$g_{a,b}(t) = \frac{1}{\sqrt{a}} g\left(\frac{t-b}{a}\right) \tag{2.19}$$

where $a > 0$ is the *dilation* parameter and b is the *translation* parameter. These parameters are typically restricted to some discrete sublattice with steps a_0 and b_0, giving

$$g_{m,n}(t) = \frac{1}{\sqrt{a_0^m}} g\left(\frac{t - nb_0}{a_0^m}\right) \tag{2.20}$$

which leads to a definition of the *discrete wavelet transform* for some signal $v(t)$

$$U(m,n) = \frac{1}{\sqrt{a_0^m}} \int_{-\infty}^{\infty} g^*\left(\frac{t - nb_0}{a_0^m}\right) v(t)dt \tag{2.21}$$

There is great latitude in the selection of a basis function, $g(t)$, and this is somewhat dependent on the application. Any well behaved, real or complex function may be used, as long as it satisfies

$$\int \frac{|G(\omega)|^2}{\omega} d\omega < \infty \tag{2.22}$$

where $G(\omega)$ is the FT of $g(t)$. Recent work has concentrated on defining wavelets which form an orthogonal basis [AS87] and those which feature compact support [Dau88a].

Various forms of the wavelet transform have been used extensively in both image [Mal89] and audio analysis [KMMG87, KM88]. Work on audio has been concentrated upon by the Marseilles group [KMMG87], and uses Gaussian wavelets of the form

$$g(t) \approx Ke^{-t^2/2}e^{j\omega_0 t} \tag{2.23}$$

where ω_0 is the characteristic frequency. Clearly this is a close relative of the basis functions of the Gabor representation (eqn. 2.15). The wavelet lattice is shown in Figure 2.1. It can be seen that varying the dilation parameter corresponds to scaling (hence multiscale) of the basis function and that this changes both the characteristic frequency of each wavelet as well as the temporal spacing between them. Note that the number of cycles in the wavelets remains constant with frequency

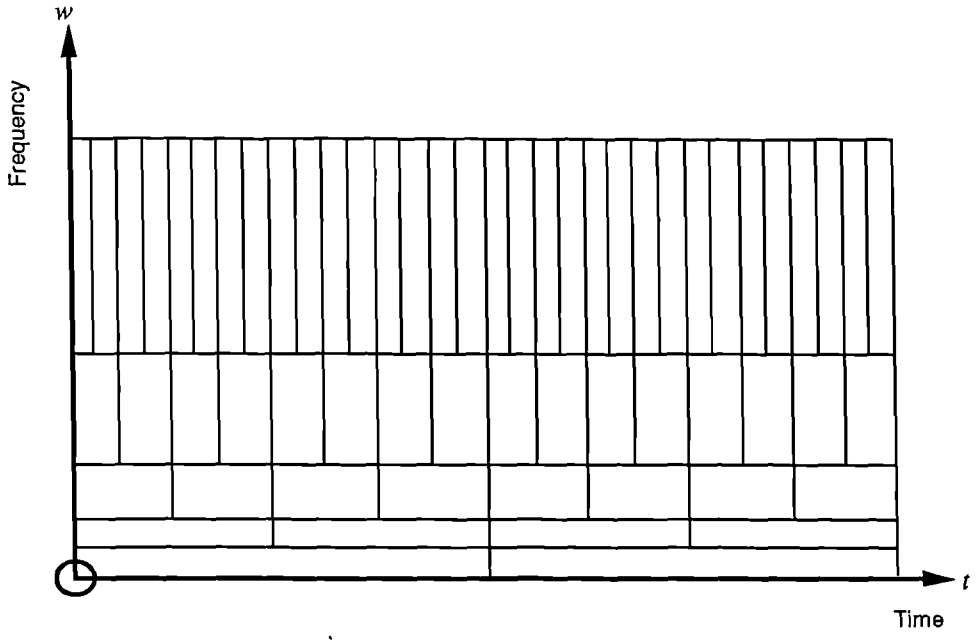

Figure 2.1: Wavelet Transform Lattice

giving a logarithmic frequency scale. The translation parameter corresponds directly to time in this application.

The transformed signal is represented over a time-dilation plane and this is similar to the time-frequency plane with the restriction that the time-frequency resolution achieved is a function of frequency. The representation has high frequency resolution at low frequencies which decreases, with a corresponding increase in temporal resolution, for higher frequencies. The range of scales used means that the representation as a whole achieves time-frequency resolutions in excess of those obtainable by using a fixed window size, but that high time *and* frequency resolution is not achieved simultaneously at any point. Analogies have been drawn between such forms of the wavelet transform and the analysis performed by the human auditory system [KMMG87], but is not clear whether this structure is advantageous for music analysis. These issues are expanded upon in the following section.

2.3 The Multiresolution Fourier Transform

All of the representations discussed have inherent advantages and disadvantages. In order to design a combined signal representation to overcome their limitations, it is necessary to identify what features would be most desirable in such a scheme.

2.3.1 Linearity and Invertibility

As described above, linearity is a highly desirable attribute of a signal transform, particularly when the signal is composed of many parts. Linearity in the signal representation greatly simplifies the analysis algorithms which operate on it; for instance, it enables linear filtering operations to be defined in terms of the transformed signal [Por80] and allows the transform to be simply interpreted. If it is required that the signal be resynthesised, a common feature of musical systems and a necessary component of coding systems, then the transform should be readily invertible. Additionally, if the transform is to be of practical use, then its definition should lead to an implementation which is computationally efficient whilst not compromising the properties of the transform.

2.3.2 Scale

Possibly the most important property of a combined representation is the time-frequency resolution achieved, as this ultimately determines the performance parameters of any analysis system built upon it. Many workers, e.g. [Moo75, Wat86, Ser89], have noted difficulties in choosing window sizes to use with linear transforms. Often such decisions must be made by trial and error, the final outcome often depending on the specific signal being analysed. The basic problem is to select a window size such that the signal features become isolated from one another, allowing their parameters to be calculated. When two or more such features lie under the same analysis window then interference will occur [WK88], making it impossible to interpret the transform coefficients in that region.

Difficulties in the selection of a single analysis window can be accounted for by observing that to achieve results comparable to the human auditory system for an arbitrary audio signal requires simultaneous time and frequency resolutions far in excess of the bounds of the uncertainty principle, i.e.

$$\Delta t \Delta f \ll 1 \tag{2.24}$$

indicating that such a window cannot exist. Multiscale techniques go some way to improving this situation by taking advantage of the observation that it may not be necessary to achieve such high resolutions simultaneously for all times and frequencies. The logarithmic frequency axis provided is similar to the human perception of pitch, but it is not clear from the literature that the ear's frequency discrimination varies in a similar manner, especially for low frequencies [Nor70]. Thus, such a form is of questionable utility for the analysis of polyphonic signals; it is true that the fundamental frequencies of notes are spaced logarithmically across frequency but the notes' partials lie at integer multiples of these frequencies resulting in many closely spaced harmonics at high frequencies. An accurate analysis of such signals requires high frequency resolution over a wide range of frequencies in order to isolate all of the signal components with separate analysis windows. Unfortunately these windows will then have long durations, increasing the likelihood that two or more temporally separated features will lie within that window. The problem here is in identifying a *natural* scale for the signal features being identified: knowledge of the appropriate scale will lead to a suitable choice of analysis window. The problem is that such knowledge is unavailable prior to analysis. A solution to this dilemma, which forms the basis of this work, is as follows. Since, for a general analysis system, it is impossible to choose a suitable window size without knowing the signal's characteristics, it will be necessary to postpone such a decision until the analysis is at least partly complete and so a *range* of window sizes must be available in the signal transform. In contrast, the wavelet transform features a range of window scales, but does not offer any *choice* of scale and so the problem of adapting the scale to the data still remains.

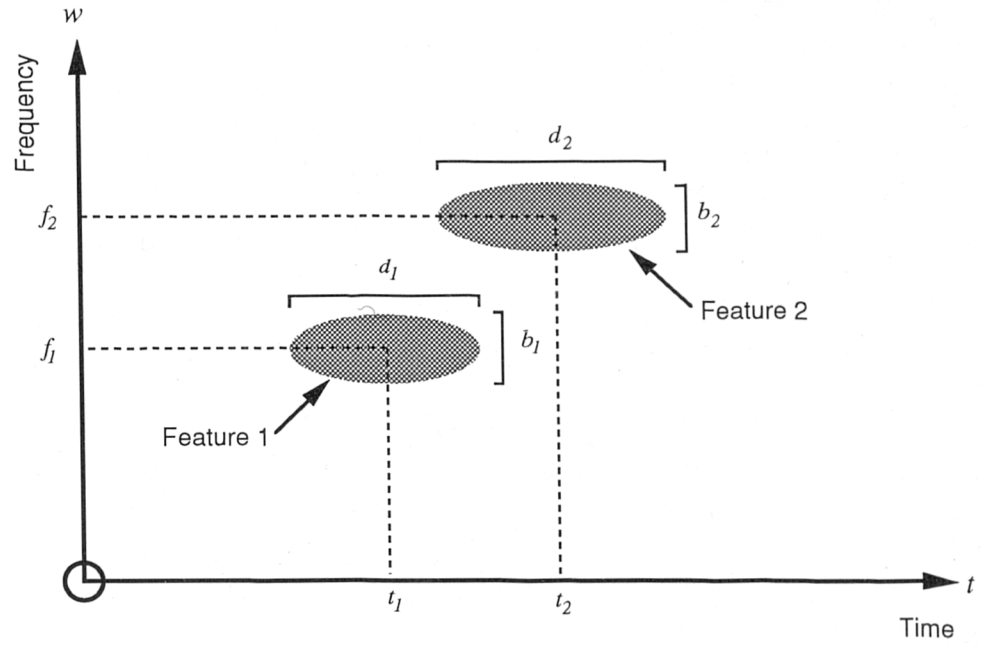

Figure 2.2: Idealised representation of a two-feature signal

What is scale?

The term 'scale' has seen much use in recent years by the signal processing community primarily as a result of the increase in interest in multiscale (wavelet) techniques. In these representations the scale of the analysis windows varies with frequency and scale is often used simply as an analogue of frequency.

By contrast in this work the term 'scale' is used independently of time or frequency suggesting that it must have an axis of its own, along which there are many representations of the time-frequency plane with differing time and frequency resolutions. It should then be possible to select the representation scale for *each* point on the time-frequency plane, whatever its position. Unlike wavelet transforms, this structure allows analysis algorithms to operate in a 'scale space' independent of time and frequency.

The effect of scale

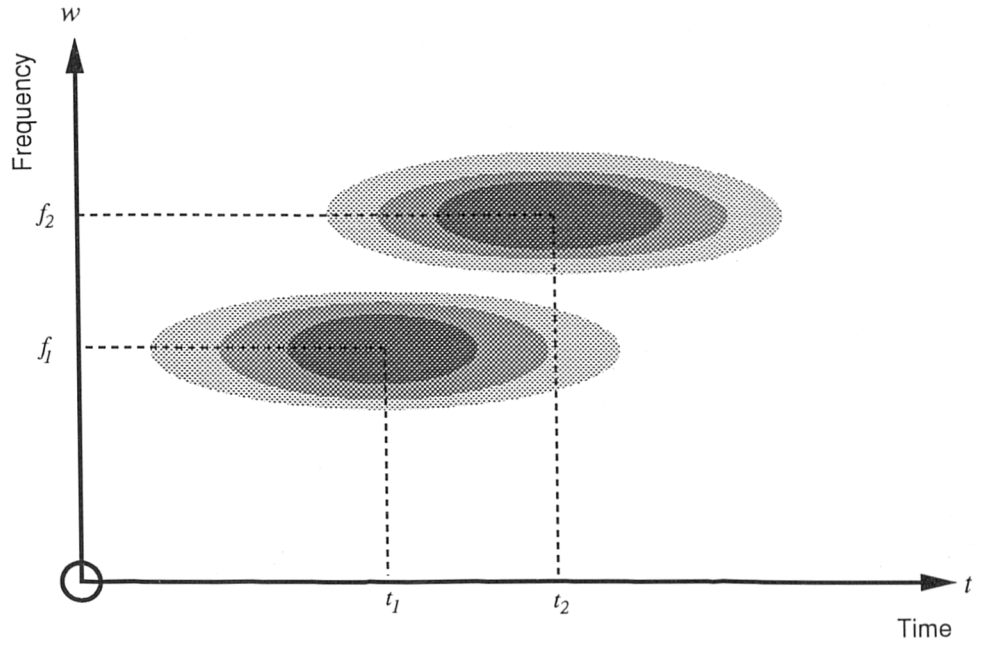

Figure 2.3: Finite resolution representation

Consider the idealised representation of the continuous time-frequency plane in Figure 2.2. The signal shown contains two features, each of which is localised in both time and frequency with the durations and bandwidths shown. The features are also distinct: there is no overlap between the areas they occupy. This representation is idealised in the sense that it does not take into account the restriction on simultaneous time-frequency resolution imposed by the uncertainty principle, it assumes arbitrary resolution.

A realisable representation of this signal will have limited resolution. An interpretation of this is that each point in the representation is associated not only with a time and frequency but also a corresponding pair of uncertainty values, determined by the time-frequency distribution of the analysis window. The implication is that when the separation of two such points is not greater than their uncertainties then their values are not independent of each other. This effect will appear as a spreading or blurring of the features. Figure 2.3 shows how a limited resolution representation of the two feature signal may appear: the time-frequency regions spread according to the choice of

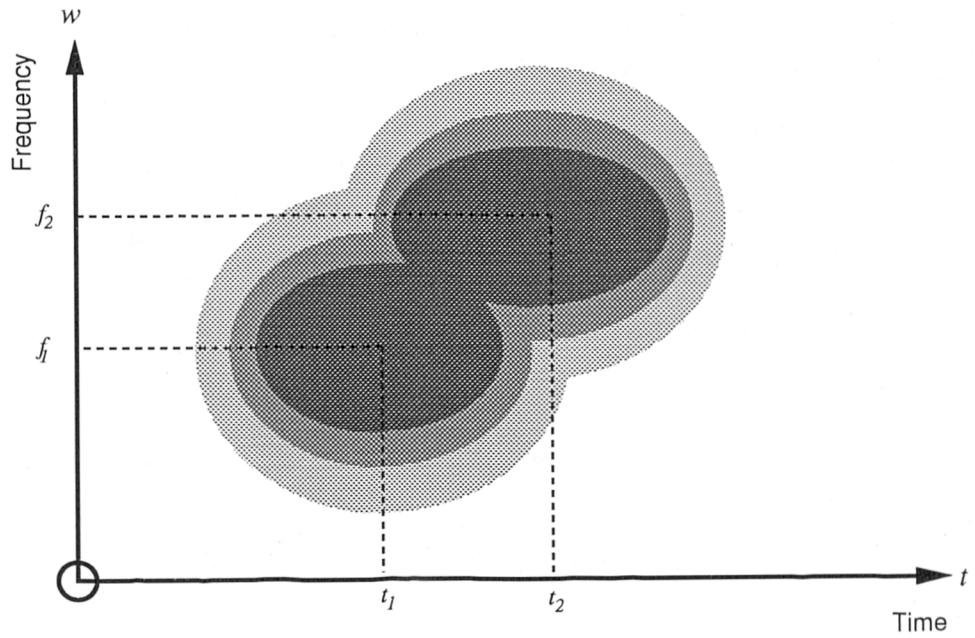

Figure 2.4: Inappropriate representation

analysis window. This will reduce the accuracy with which the feature parameters can be measured. Figure 2.4 shows a different case: at this scale the feature regions are no longer distinct resulting in interference between the features [WK88]. It may not be possible to identify and parameterise the two features using this representation. Indeed they could be classified as a single feature. Given that the signal model defines two features in the signal, such a classification would be incorrect and this leads to the observation that this scale is somehow *inappropriate* for the analysis of such a signal.

Finding the correct scale

How can an appropriate scale be determined? Clearly the separation of the two features plays an important role: the further they are apart the more likely they are to be correctly identified. This relationship between scale and feature separation is best explored by considering some simple examples in which aspects of time and frequency can be isolated.

Figure 2.5: $|V_a(t,\omega)|$ (Infinite resolution)

A Single Feature. A simple example signal consists of a single complex sinusoid

$$v_a(t) = e^{j\omega_a t} \tag{2.25}$$

where ω_a is the constant frequency of v_a. This signal is infinite in time, having no beginning or end. The idealised representation of this signal is shown in Figure 2.5, its energy is entirely concentrated in frequency at ω_a and distributed evenly across all time. The STFT representation of v_a can be found

$$\begin{aligned} V_a(t,\omega) &= \sum_\tau w(\tau - t)e^{-j\omega\tau}v_a(\tau) \\ &= \sum_\tau w(\tau - t)e^{-j\omega\tau}e^{j\omega_a\tau} \\ &= \sum_\tau w(\tau)e^{-j(\omega-\omega_a)(\tau+t)} \\ &= W(\omega - \omega_a)e^{-j(\omega-\omega_a)t} \end{aligned} \tag{2.26}$$

where $w(t)$ is the analysis window and $W(\omega)$ is its FT. The local magnitude spectrum at any time is thus the FT of the analysis window shifted in frequency to ω_a. The limit on frequency resolution due to the width of the analysis window causes the signal's energy to appear spread in frequency. An example is shown in Figure 2.6 for the window

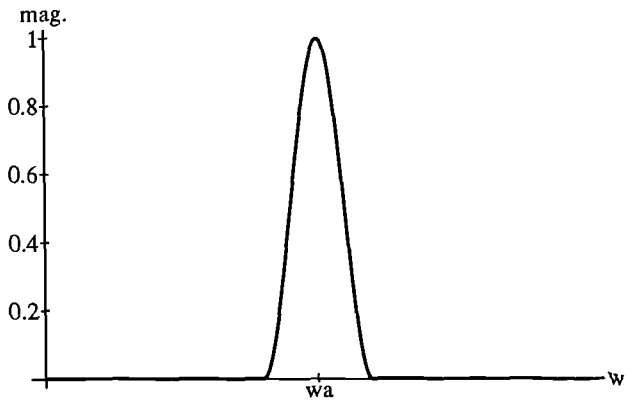

Figure 2.6: $|V_a(t_0, \omega)|$ (Scale 1)

$$W(\omega) \approx \begin{cases} 1 + \cos(\frac{\pi}{\Omega}\omega) & |\omega| \leq \Omega \\ 0 & \text{else} \end{cases} \tag{2.27}$$

where Ω controls the window bandwidth. The relationship between such a representation and the idealised model is still clear, however. There is still a single stationary band of energy running across time. An appropriate choice of analysis windows and sampling intervals will lead to accurate estimation of the single signal parameter, its frequency. Suppose now that the scale of the representation is changed: the altered window function is more concentrated in time and correspondingly wider in frequency giving increased temporal resolution. The resulting local magnitude spectra of the representation are naturally more dispersed (Figure 2.7) but the form is the same as before: a single peak. Provided that the sampling intervals of this new representation are modified appropriately then the signal frequency may still be obtained without ambiguity. It seems that this signal is accurately represented *irrespective* of the analysis scale used.

Two Features. Consider now a signal comprising two complex exponentials

$$v_b(t) = e^{j\omega_1 t} + e^{j\omega_2 t} \tag{2.28}$$

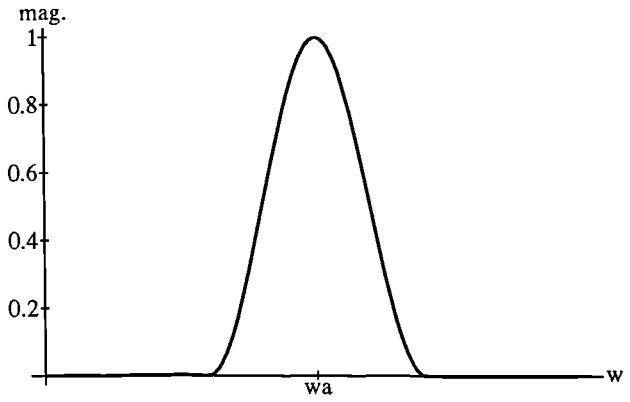

Figure 2.7: $|V_a(t_0, \omega)|$ (Scale 2)

with frequencies ω_1 and ω_2. The frequencies and particularly their separation, $\omega_2 - \omega_1$, could be chosen so that a listener would be able to hear two distinct tones when presented with two sinusoids at these frequencies, for example. In such a case, the desired segmentation of the signal will isolate the two components and estimate their frequencies. Figure 2.8 shows an idealised time-frequency representation of this signal, two distinct bands of energy which is easily related to the listener's perceptions. Forming windowed transforms of this signal at the two scales used before gives the pair of local magnitude spectra in Figures 2.9 and 2.10. Given that the representation is linear these may be obtained from Equation (2.26), giving

$$V_a(t, \omega) = W(\omega - \omega_1)e^{-j(\omega-\omega_1)t} + W(\omega - \omega_2)e^{-j(\omega-\omega_2)t} \tag{2.29}$$

At the scale with higher frequency resolution (Scale 1) there are two distinct peaks at the frequencies of the signal components. Assuming that the analysis windows are exactly bandlimited then the signal is perfectly segmented, the magnitudes of the peaks are independent of each other depending solely on the magnitude of the corresponding component. An alternative interpretation of this is given by observing that the width of the analysis window in frequency is smaller than the frequency

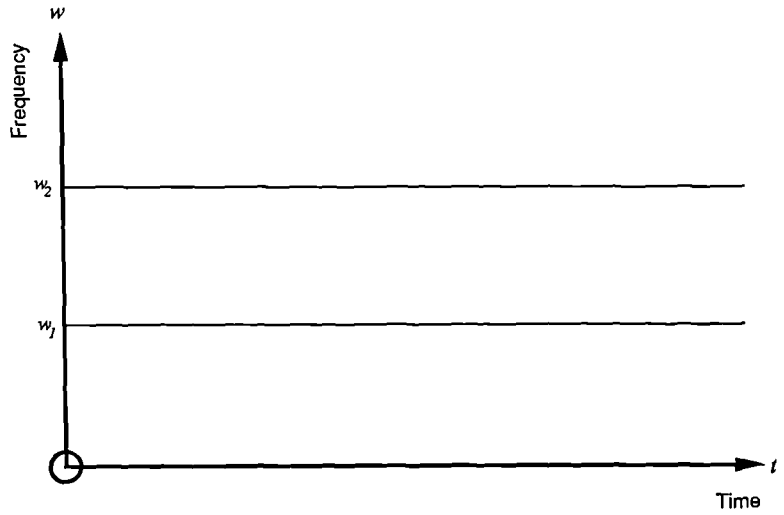

Figure 2.8: $|V_b(t,\omega)|$ (Infinite resolution)

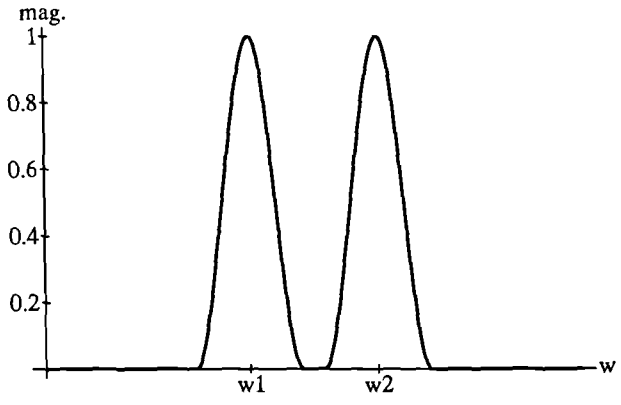

Figure 2.9: $|V_b(t_0,\omega)|$ (Scale 1)

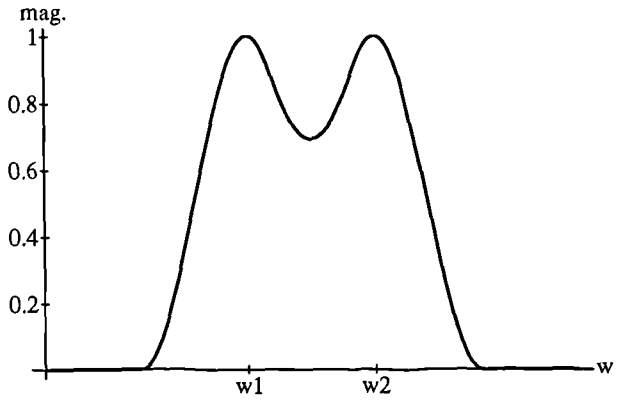

Figure 2.10: $|V_b(t_0, \omega)|$ (Scale 2)

separation of the two features and so it is impossible for more than one feature to 'fall' under the window whatever its centre frequency. However, at the scale with lower frequency resolution (Scale 2) this is no longer true, the bandwidth of the analysis window is now greater than the frequency separation of the features so the two components are less well separated. At this scale the representation no longer conforms to the signal model (eqn. 2.28). A feature detection algorithm using that signal model may fail when presented with such data. The amount of interference between features which can be tolerated by a detector is clearly dependent on the algorithm it uses, but there will always be some amount above which correct separation cannot be achieved. Such scales can be considered inappropriate for the required segmentation of the signal and this implies the existence of a scale bound based on the frequency separation of the features, in terms of the minimum frequency resolution required. It is not the case, however, that it will be possible to separate any pair of features simply by choosing a scale within some bound. A more accurate model can be formed in terms of the probability of separating some pair of features which will increase with their separation.

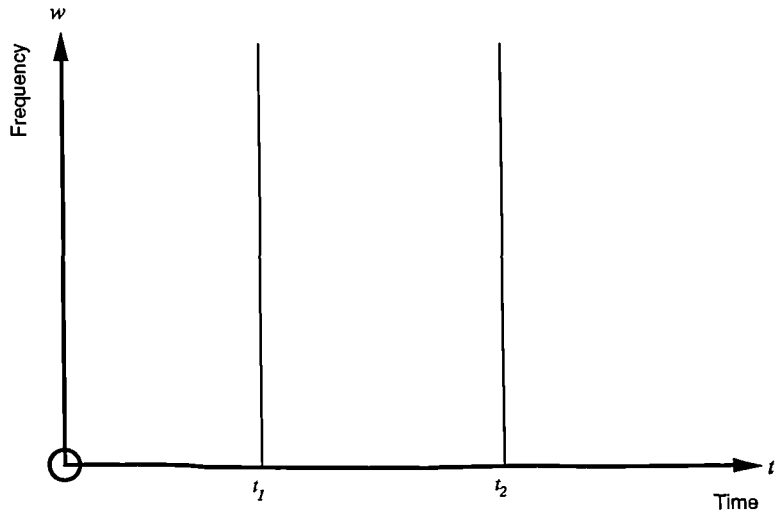

Figure 2.11: $|V_c(t,\omega)|$ (Infinite resolution)

Scale Restriction 1 *The frequency separation of some pair of features will restrict the choice of analysis scale by defining a frequency resolution below which the probability of successful detection, with a given window, is unacceptably low.*

The same effect can be observed in the time domain by considering a signal

$$v_c(t) = \delta(t - t_1) + \delta(t - t_2) \tag{2.30}$$

composed of two impulses at times t_1 and t_2. Each impulse is a purely temporal feature fully parameterised by a single time value. Assuming that the signal analysis is to determine the times of these impulses in the signal then a similar argument to that above can be followed to restrict the range of applicable analysis scales. Each impulse is broadband in frequency and highly concentrated in time, this can be seen in the idealised time-frequency representation in Figure 2.11. The correspondence with Figure 2.8 is clear: the axes are transposed. Sections parallel to the time axis of an STFT of this signal will have the same form as the spectra in figures 2.9 and 2.10, leading to the conclusion that to distinguish the two impulses the representation used must employ an analysis window with a duration shorter than the temporal separation of the impulses (assuming the window

is now finite in time).

Scale Restriction 2 *The temporal separation of two features will restrict the choice of analysis scale by defining a temporal resolution below which the probability of successful detection, with a given window, is unacceptably low.*

Feature Localisation. The signals discussed in the previous discussion were each localised in only one domain. Clearly such features are idealised and do not occur in natural signals. This work seeks to describe musical signals in terms of a set of features localised, to some degree, in both time and frequency. The localised nature of these features causes them to be 'spread out', making it possible to describe them in terms of their widths (duration and bandwidth) in these domains. A common requirement of feature detection algorithms is that the feature to be detected should fall entirely under at least one of the analysis windows in one or both dimensions. The idealised features examined above are unrealistic: local structure variation is common in natural signals and so the detection algorithms require an analysis window with sufficient width to span the feature. Such local structure decreases the certainty with which the feature's parameters can be determined and gives rise to a further two scale restrictions required to guarantee an appropriate analysis scale.

Scale Restriction 3 *For a given detection strategy, the local bandwidth of the target feature restricts the highest frequency resolution which may be used to detect it reliably.*

Scale Restriction 4 *For a given detection strategy, the local duration of the target feature restricts the highest temporal resolution which may be used to detect it reliably.*

Restrictions have now been defined for the upper and lower time and frequency resolutions required for the accurate detection of some feature. The lower restrictions depend on its separation from neighbouring features in the appropriate domain while the upper restrictions are related to the local structure of that feature. For a given window, these restrictions define a range of scales at

which it may be applied in order to reliably identify some target feature; indeed it may be said that they define the 'natural' scale of a feature.

The General Case Returning now to the more general segmentation problem for the signal shown in Figure 2.2, it can be seen that all four restrictions can be determined for each feature e.g. Feature 1:

$$\Omega < f_2 - f_1 \tag{2.31}$$

$$\Gamma < t_2 - t_1 \tag{2.32}$$

$$\Omega > b_1 \tag{2.33}$$

$$\Gamma > d_1 \tag{2.34}$$

Unfortunately the situation is not so simple because the maximum time and frequency resolutions available simultaneously are also bounded by the uncertainty principle (eqn. 2.16) i.e. the four scale restrictions are not independent.

The second pair of restrictions is independent of the signal context, depending only on the extent, in time and frequency, of the feature in question and this can be determined by interactive analysis of various examples of the feature represented over a range of scales. Signal context does, however, determine the first pair of restrictions. Knowledge of the nearest neighbouring features in time and frequency is required before suitable scales can be selected which isolate the feature under investigation. However, the determination of the signal context requires the parameterisation of the neighbouring features which also requires knowledge of their context. This dilemma can be resolved at least partly by careful ordering of the analysis steps, see Chapter 6.

Situations may arise, of course, when many features are in close proximity, in which no appropriate analysis scale n can be selected to satisfy the requirements of a detection strategy based on single feature localisation. In these cases interference between features is inevitable and it will be difficult to extract meaningful features from the data unless the detectors can be made insensitive to such interference. Alternatively, detectors could be devised to recognise more than

one feature from a local data set. Such a scheme for determining the frequencies of two unresolved partials within a duet signal has been described by [Mah90] but it is too specialised for a general system. Investigations carried out during the preparation of this work have not revealed any suitable algorithms which reliably implement such multiple-feature detectors for natural signals. The large quantity of local structure variation in such signals (e.g. the beating of coincident partials) introduces a high degree of ambiguity to the data, giving poor results for the simultaneous detection of just two features.

2.3.3 Invariance.

It was mentioned in the introduction that it is desirable for an interactive analysis-synthesis system that the signal representation should be intuitively interpretable by a human observer, additionally this quality simplifies the development of analysis algorithms. Linearity contributes to this, but it is greatly enhanced if the transform shares invariances with the signal features being detected. To illustrate: a note may occur at any point in time; if it is delayed by some amount then none of its perceptual properties (other than onset time) is changed - it is time-shift invariant. It is desirable that the representation of such a note is also unchanged, apart from a time shift. Similarly, if a note's duration is altered or it is transposed in pitch then its representation should reflect this change while remaining unmodified in all other respects. Note that the structure of the wavelet transform gives rise to time-shift invariance but not frequency-shift invariance, so that a note and its transpose would be represented by differing numbers of coefficients. Time and frequency invariances can be satisfied by the use of a regular tessellation of the time-frequency plane i.e. regular sampling intervals of both time and frequency, as is used in the STFT.

2.3.4 Representation Requirements.

To summarise, the desirable properties of a signal representation for musical signal analysis are:

1. Linearity.

2. Invertibility.

3. A choice of resolution at each point on the time-frequency plane.

4. Invariances which correspond to a listener's perceptions.

5. Simple and efficient implementation.

In order to satisfy these demands the approach taken in this work is to use a transform based on the STFT, but which provides multiple representations of the signal, each of which has a different time-frequency resolution. The transform is referred to as the *multiresolution Fourier transform* (MFT) and is a one dimensional form of a scheme recently developed for image analysis [Cal89]. An alternative description of the MFT, including the 2-d form, can be found in that work and [WCPon].

2.3.5 MFT Definition and Properties

The MFT is best introduced by considering again the continuous STFT, but now introducing a scale parameter σ such that

$$\hat{v}(t,\omega,\sigma) = \int_{-\infty}^{\infty} w_\sigma(t-\tau)v(\tau)e^{-j\omega\tau}d\tau \tag{2.35}$$

The scale parameter affects the size of the analysis window which is related to a basic analysis window $w(t)$ via

$$w_\sigma(t) = \sigma^{\frac{1}{2}} w(t\sigma) \tag{2.36}$$

Thus as σ decreases, the duration of the analysis window increases, allowing it a greater concentration of energy in the frequency domain. As $\sigma \to 0$ then the window becomes infinitely long and Equation (2.35) reduces to the continuous Fourier transform.

Using this structure the MFT can be considered to be a superset of both the wavelet and STFT representations. This may be seen by observing the invariances of the MFT's set of analysis vectors,

$\gamma_{t,\omega,\sigma}(\tau)$, to various transformations. The analysis vectors are simply the appropriate time and frequency shifts of the analysis windows [Dau88a]

$$\gamma_{t,\omega,\sigma}(\tau) = w_\sigma(\tau - t)e^{-j\omega\tau} \tag{2.37}$$

The delaying of such a vector by time δ_t gives

$$w_\sigma(\tau + \delta_t - t)e^{-j\omega(\tau+\delta_t)} = \gamma_{t+\delta_t,\omega,\sigma}(\tau)e^{-j\omega\delta_t} \tag{2.38}$$

a different analysis vector with an appropriate phase shift. Similarly the frequency shift of an analysis vector by δ_f transforms to another analysis vector.

$$w_\sigma(\tau - t)e^{-j(\omega+\delta_f)\tau} = \gamma_{t,(\omega+\delta f),\sigma}(\tau) \tag{2.39}$$

These invariances are similar to those of the STFT. Additionally the MFT's analysis vectors are invariant to a dilation by a factor δ_σ. Using Equation (2.36), such a dilation gives

$$w_\sigma(\delta_\sigma\tau - t)e^{-j\omega\tau} = \gamma_{t/\delta_\sigma,\delta_\sigma\omega,\delta_\sigma\sigma}(\tau) \tag{2.40}$$

which corresponds directly to the same operation in the wavelet transform.

The discrete form of the MFT can now be defined. The overall structure is shown in Figure 2.12. The representation is indexed by three independent parameters: time, frequency and time-frequency resolution. The resolution parameter n selects one of a number of transform *levels*, each of which is a complete invertible description of the signal, with a structure identical to that of the discrete STFT. Resolution varies uniformly between levels, with the lowest and highest levels being the original signal and its DFT respectively. The level n relates to the scale parameter of the continuous form via

$$\sigma = \frac{1}{\alpha^n} \tag{2.41}$$

where α is the MFT scale constant. This multiplicity of representations allows the analysis algorithms (rather than their implementer) to 'choose' the most appropriate window size to use according to the signal content, giving the ability for the system to adapt to widely varying input.

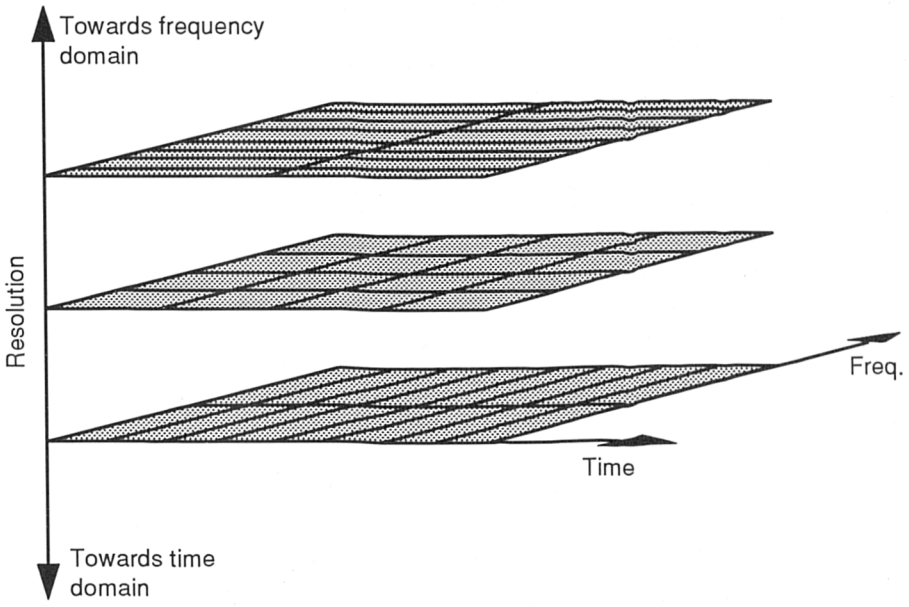

Figure 2.12: 1-d MFT structure

For a finite input signal sequence $\{x_i\} : 0 \leq i < M$ the transform coefficients are given by

$$\hat{x}_{ik}(n) = \sum_{l=0}^{M-1} g_{(l-i\Gamma(n))}(n) x_l e^{-j\frac{2\pi}{M}\Omega(n)kl} \tag{2.42}$$

where $\mathrm{g}(n)$ is the analysis window for level n, $\Gamma(n)$ and $\Omega(n)$ are the time and frequency sampling intervals and all index arithmetic is calculated modulo M. The indices i and k on each level select coefficients from a regular lattice covering the time-frequency plane. There are different numbers of coefficients on each axis for each level, with

$$0 \leq i < N_i(n) \tag{2.43}$$

and

$$0 \leq k < N_k(n) \tag{2.44}$$

For each level to be a complete description of the original sequence, the sampling theorem states that there must be at least the same number of coefficients per level as there were samples in the sequence i.e.

$$N_i(n) N_k(n) \geq M \tag{2.45}$$

Choosing the scale constant, $\rho = 2$, and constraining the signal sequence length M to be a power of two, leads both to efficient computation via the FFT and an obvious choice for the number of sample points to satisfy the equality in Equation (2.45)

$$N_i(n) = 2^{N-n} \;\; N_k(n) = 2^n \;\; M = 2^N \;\; 0 < n < N \tag{2.46}$$

and so, for regular sampling, the time and frequency sampling intervals,

$$\Gamma(n) = N_k(n) \;\; \Omega(n) = N_i(n) \tag{2.47}$$

Thus each MFT coefficient $\hat{x}_{ik}(n)$ represents a rectangular region or 'cell' of the $M \times M$ time-frequency plane, of size $\Gamma(n) \times \Omega(n)$ and position $i\Gamma(n), k\Omega(n)$. These cells are arranged as a uniform time-frequency plane tessellation. Note that each coefficient on level $n+1$ represents a region of size $2\Gamma(n) \times \frac{1}{2}\Omega(n)$, giving a doubling ($\alpha = 2$) of frequency resolution and corresponding reduction of temporal resolution; consequently the coefficient pair

$$\left[\hat{x}_{ik}(n), \hat{x}_{(i+1)k}(n)\right] \tag{2.48}$$

represents the same area of the time-frequency plane as the pair

$$\left[\hat{x}_{(i/2)(2k)}(n+1), \hat{x}_{(i/2)(2k+1)}(n+1)\right] \tag{2.49}$$

on the level above.

The MFT definition leads to the requirements for the analysis windows $g(n)$, clearly they must be localised in both time and frequency, and ideally should be zero outside a time-frequency region of size $\Gamma(n) \times \Omega(n)$.

$$g_i(n) = 0 \quad \text{if} \quad i < 0 \;\; i \geq \Gamma(n) \tag{2.50}$$

$$\hat{g}_i(n) = 0 \quad \text{if} \quad i < 0 \;\; i \geq \Omega(n) \tag{2.51}$$

However, it has been shown that no such function exists [Pap77] and so this ideal is compromised somewhat by using a window with finite bandwidth $\Omega(n)$ but which is concentrated, as much as is allowable, in the time interval $\Gamma(n)$.

These requirements suggest the use of the class of finite prolate spheroidal sequences 'FPSSs', which have been fully described in [PSL61] and [WS87a]. The definition of these sequences is best put in terms of a solution to an eigenvector problem, using linear operator notation

$$\mathbf{B}(\Omega(n))\mathbf{T}(\Gamma(n))\mathbf{g}(n) = \lambda_0 \mathbf{g}(n) \tag{2.52}$$

where $\mathbf{T}(\Gamma)$ is the index limiting operator, an $M \times M$ matrix with elements

$$T_{ik}(\Gamma) = \begin{cases} \delta_{ik} & |k| < \Gamma/2 \\ 0 & \text{else} \end{cases} \tag{2.53}$$

The DFT operator $\mathbf{F}$ can be similarly defined with elements

$$F_{ik} = \frac{1}{\sqrt{M}} e^{-j\frac{2\pi}{M}ik} \tag{2.54}$$

The operator $\mathbf{B}(\Omega)$ in Equation (2.52) is the bandlimiting operator, the frequency domain counterpart of $\mathbf{T}$, which can be defined as

$$\mathbf{B}(\Omega) = \mathbf{F}^*\mathbf{T}(\Omega)\mathbf{F} \tag{2.55}$$

where $\mathbf{F}^*$ is the adjoint of $\mathbf{F}$ and is thus the inverse DFT operator. The scalar λ_0 is the largest eigenvalue of the combined operator $\mathbf{B}(\Gamma(n))\mathbf{T}(\Omega(n))$ and $\mathbf{g}(n)$ is the associated eigenvector. In other words an FPSS is a sequence which is unchanged, apart from a linear factor, by the application of time-truncation and bandlimiting operations. The resulting sequences do satisfy the requirements of the MFT analysis windows: it can easily be seen from the definition of the combined operator that these sequences are exactly bandlimited in the interval $\Omega(n)$, and it has been shown that they have their energy optimally concentrated in the time interval $\Gamma(n)$ [WS87a]. The use of bandlimited analysis windows allows efficient computation of the MFT using the FFT algorithm.

Extending this linear operator notation, each level of the MFT can be defined as

$$\hat{\mathbf{x}}(n) = \mathbf{F}(n)\mathbf{x} \tag{2.56}$$

$\mathbf{F}(n)$ is the nth level MFT operator. The parameters of the MFT levels in Equation (2.46) lead to the observation that the lowest level is simply

$$\mathbf{F}(0) = \mathbf{I} \tag{2.57}$$

the identity operator. At the other extreme, the highest level is the DFT of the original sequence

$$\mathbf{F}(N) = \mathbf{F} \tag{2.58}$$

2.3.6 Oversampled Transform

The above discussion is based on an MFT where the number of coefficients on each level is the minimum required to satisfy the sampling theorem. However, Equation (2.45) suggests that alternative structures are possible in which a certain degree of oversampling is incorporated. Discussion of the issues behind this type of MFT is postponed until the next chapter, where they may be considered alongside the implementational details of the transform.

2.3.7 Inverse Transform

The overcompleteness of the MFT leads to a number of possibilities when defining an inverse transform. Each level of the MFT contains enough information to reconstruct the original signal exactly. Thus an inversion operator, $\mathbf{F}^{-1}(n)$, may be defined for each level, such that

$$\mathbf{x} = \mathbf{F}^{-1}(n)\hat{\mathbf{x}}(n) \tag{2.59}$$

The observation that, since the MFT analysis windows are bandlimited to the frequency domain sampling interval, then the rows of the transform are linearly independent, leads to a definition of the inverse transform's synthesis windows $\mathbf{g}^{-1}(n)$ via the frequency domain relationship [Cal89]

$$\hat{g}_k^{-1} = \begin{cases} \frac{1}{\hat{g}_k(n)\Omega(n)} & |k| < \Omega(n)/2 \\ 0 & \text{else} \end{cases} \tag{2.60}$$

where $\hat{\mathrm{g}}(n)$ and $\hat{\mathrm{g}}^{-1}(n)$ are the Fourier transforms of $\mathrm{g}(n)$ and $\mathrm{g}^{-1}(n)$ respectively. In other words the synthesis window has the inverse frequency response of the corresponding analysis window within the interval of its support.

An alternative strategy is to devise some sort of inverse operator which uses a set of coefficients selected from more than one level. Clearly there are many possible schemes for the selection process and the choice may be dependent on the application. Some examples have been considered for image reconstruction in [Cal89], but this theme is not pursued further here.

2.3.8 MFT Interpretation

Interpretation of the MFT coefficients is analogous to that for the STFT. The analysis window is localised in both time and frequency, the magnitude of the coefficient thus represents the amount of energy present in a particular region of the time-frequency plane over which the signal is expanded. The uniform time-frequency plane tessellation of each MFT level leads directly to the local spectrum and filter bank interpretations suggested for the STFT [Por81].

An MFT level can be considered an implementation of a filterbank structure by observing that each row in Equation (2.42) can be interpreted as a subsampled convolution of an analysis filter $\mathrm{g}(n)$ with a frequency shifted version of the signal sequence $x_i e^{-j\omega_k i}$, where $\omega_k = \frac{2\pi}{M}\Omega(n)k$. The filters in the bank will have an impulse response similar to the analysis window in the time domain [Por81].

Alternatively, by selecting one column $i = i_0$ of a level, Equation (2.42) can be written as

$$\hat{x}_{i_0 k}(n) = \sum_{l=0}^{M-1} x_l(i_0) e^{-j\omega_k l} \tag{2.61}$$

which is the DFT of the windowed sequence

$$x_l(i) = g_{(i\Gamma(n)-l)}(n) x_l \tag{2.62}$$

Each column of a level can thus be interpreted as a local Fourier spectrum of the input signal viewed

through the analysis window.

The importance of the use of phase derivatives with respect to both time and frequency in feature identification is discussed in chapters 4, 5 and 6. These partial derivatives must be approximated, in the case of a discrete transform, by the corresponding phase differences which may be obtained via

$$\mu_{ik}(n) = \arg(\hat{x}_{ik}(n)\hat{x}^*_{i(k-1)}) \tag{2.63}$$

$$\nu_{ik}(n) = \arg(\hat{x}_{ik}(n)\hat{x}^*_{(i-1)k}) \tag{2.64}$$

where $*$ indicates the complex conjugate. It can be seen that the desired choices of sampling intervals and analysis window would allow these values to be determined unambiguously from the MFT coefficients. However the choice of window was compromised in that it did not satisfy Equation (2.51) resulting in a certain amount of leakage between time frames. A means of reducing these errors is discussed in the next chapter.

MFT interpretations involving coefficients from more than one level are fully discussed in Chapter 6, for now it can be seen that the MFT definition leads to simple inter-level relationships such as that between coefficients in equations (2.48) and (2.49). Such simplicity encourages the development of multiresolution models and algorithms using the MFT.

2.3.9 Summary

This Chapter has reviewed several time-frequency signal representations and proposed a new representation, the MFT, designed to overcome their disadvantages.

Chapter 3

MFT Implementation and Initial Results

3.1 Introduction

The previous chapter gave a definition for the MFT and described its properties: the purpose of this chapter is to give a more practical discussion of the MFT and its application to audio analysis. The first section discusses the form of MFT best suited to this application, this is followed by a description of the actual implementation used in this work. The second half of the chapter presents examples of the MFT applied to some simple signals and discusses their properties.

3.2 Selecting the MFT Parameters

The MFT described in the previous chapter consisted of N transform levels ranging in resolution from the original signal sequence up to its DFT. It has been indicated in several places above that the kinds of signal features of interest to an analysis system are best localised at a scale some way intermediate between these two extremes, suggesting that the extreme levels may be of little use for analysis work. It then follows that this may also be true for other levels close to these extremes, which leads to the conclusion that it may not be necessary, or desirable for storage considerations,

Level	Coefficients		
	Time	Frequency	
0	65536	1	Time Sequence
1	32768	1	
2	16384	2	
3	8192	4	
4	4096	8	
5	2048	16	
6	1024	32	
7	512	64	
8	256	128	
9	128	256	
10	64	512	
11	32	1024	
12	16	2048	
13	8	4096	
14	4	8192	
15	2	16384	
16	1	32768	Fourier Transform

Table 3.1: Coefficients per MFT Levels

to generate all N levels of the MFT. Clearly, for a given application, there will be some useful range of levels, $n_l \ldots n_h$, such that

$$0 < n_l \leq n_h < N \tag{3.1}$$

The parameters which serve to characterise each level are:

1. The width of the analysis vector in the frequency domain.

2. The temporal duration of that vector within which its energy is concentrated.

3. The frequency sampling interval, i.e. the distance between the centres of adjacent frequency bins.

4. The temporal sampling interval; this is commonly referred to as the *hop-size*.

Example values for these parameters are given in tables 3.1 and 3.2 for all the levels of an MFT with $M = 2^{16}$ and a sampling rate of 48 KHz. Various workers have commented on the temporal

Level	Coefficients		
	Duration (ms.)	Bandwidth (Hz.)	
0	0.02	48000	Time Sequence
1	0.04	24000	
2	0.08	12000	
3	0.16	6000	
4	0.33	3000	
5	0.66	1500	
6	1.33	750	
7	2.67	375	
8	5.33	187.5	
9	10.7	93.75	
10	21.3	46.88	
11	42.7	23.44	
12	85.3	11.72	
13	170	5.86	
14	341	2.93	
15	683	1.46	
16	1365	0.73	Fourier Transform

Table 3.2: Coefficients per MFT Levels

resolutions desirable for audio analysis; Serra in [Ser89] uses the STFT with window sizes as small as 25 ms. and hop-sizes down to 6 ms, while Watson [Wat86] chooses a hop-size of 10 ms. The frequency resolution required depends heavily on the application; for the purposes of polyphonic music transcription we note that there is 1.64 Hz separating the lowest two notes on a piano and that this represents a fairly extreme case.

The MFT considered up to this point has minimum redundancy, each level contains just enough coefficients to be a complete description of the original signal. Section 2.3.6 introduced the idea that alternative structures are possible in which this optimality is relaxed by introducing some degree of oversampling. It has been found, both in this work and in [Cal89] that these modified forms have a number of advantages over the original definition.

Consider the analysis vector shown in Figure 3.1, which satisfies Equation (2.52) with $\Gamma = 16$ and $\Omega = 32$; the chosen bandwidth gives rise to discontinuities in the frequency response and consequently the vector has fairly large sidelobe magnitudes (see Table 3.3) in the time domain.

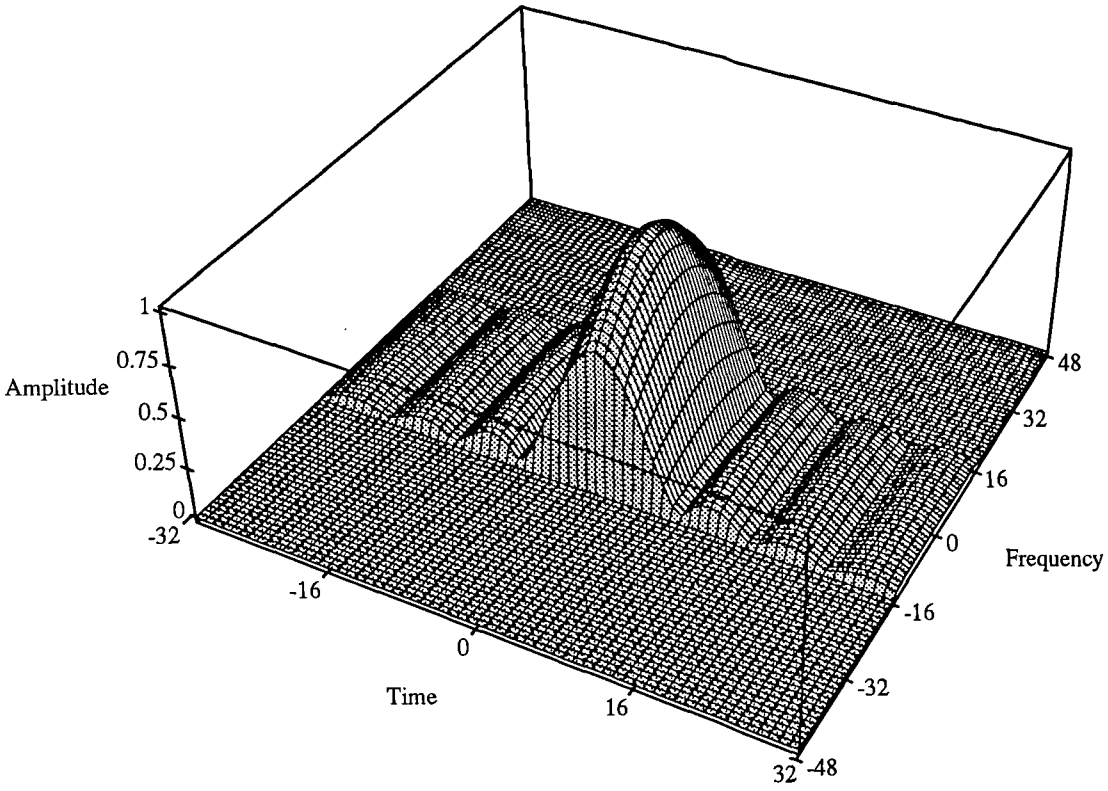

Figure 3.1: Time-Frequency plane FPSS 16 x 32

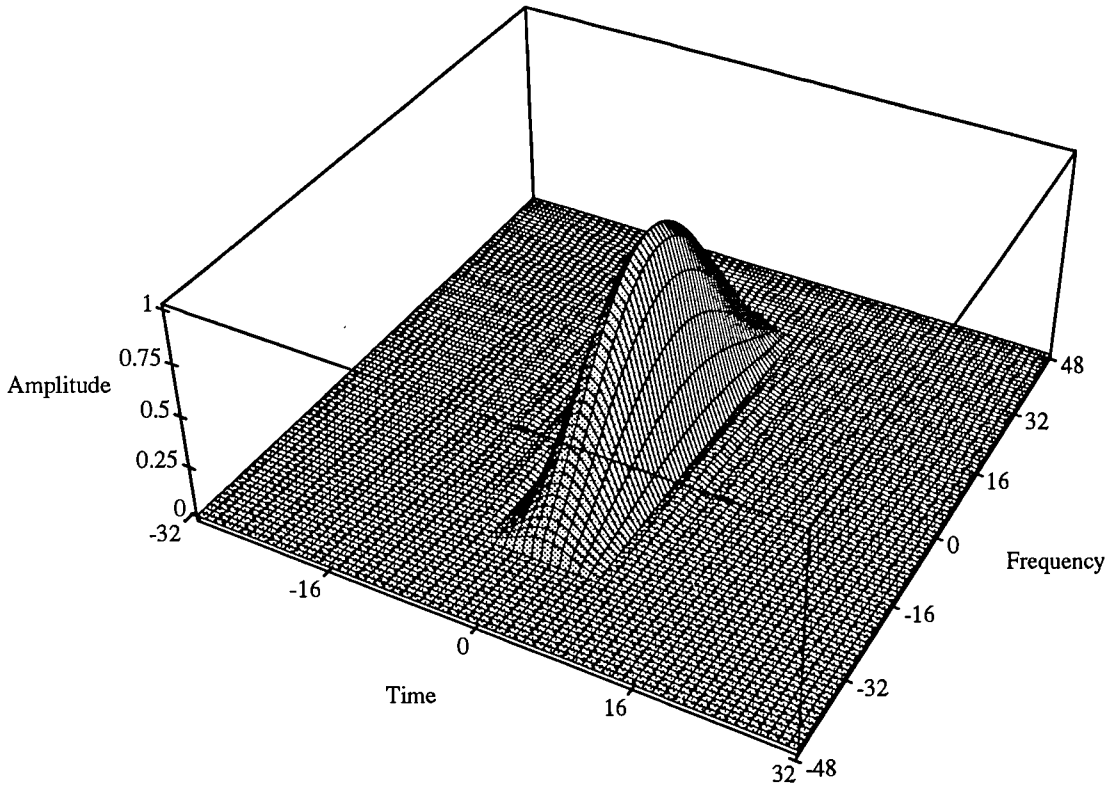

Figure 3.2: Time-Frequency plane 'relaxed' FPSS 16 x 32

Lobe	Magnitude (dB)
1	-15.6
2	-20.2
3	-23.0
4	-25.2

Table 3.3: Time Domain Sidelobe Magnitudes for FPSS 16 × 32

Lobe	Magnitude (dB)
1	-26.9
2	-32.8
3	-37.1
4	-40.0

Table 3.4: Time Domain Sidelobe Magnitudes for 'relaxed' FPSS 16 × 32

Increasing the truncation width in the frequency domain should result in a vector which is better behaved in both domains. The effect of doing this is shown in Figure 3.2 for an increase by a factor of two. The vector is smoother in the frequency domain and more localised in time, while the sidelobe magnitudes are reduced to the figures shown in Table 3.4. It is related to the original analysis vector by

$$\mathbf{g}'(n) = \mathbf{B}(2\Omega(n))\mathbf{T}(\Gamma(n))\mathbf{g}(n) \tag{3.2}$$

and has been termed a 'relaxed' FPSS owing to the relaxation of the frequency domain constraint. In order to use this modified analysis vector the sampling theorem dictates that, to retain invertibility and the phase relationships in equations (2.64) and (2.64), it is necessary to introduce a corresponding amount of temporal oversampling. The number of coefficients on each level becomes

$$N_i' = 2N_i(n) = 2^{N-n+1} \quad N_k' = N_k(n) = 2^n \quad N_i'(n)N_j'(n) = 2M \tag{3.3}$$

This modified structure has certain other advantages as well as increased temporal localisation. Figure 3.3 shows, for the original scheme, the frequency domain alignment of the shifts of the analysis vector required to generate an MFT level. The corresponding time domain relationship is shown in Figure 3.4. There is no overlap between analysis vectors in the frequency domain; this may cause 'boundary' problems when some signal feature lies very close to such a frequency

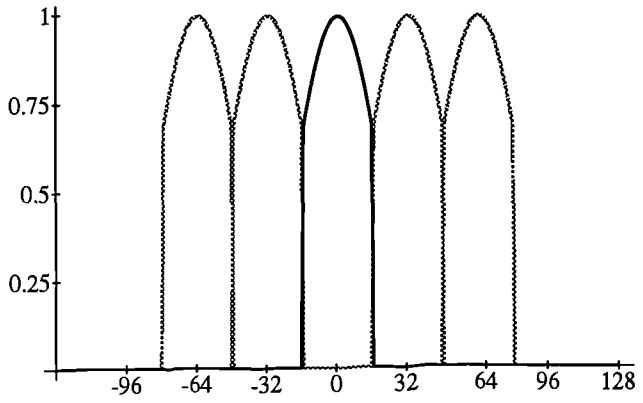

Figure 3.3: Analysis vectors (freq), $\sigma = 1$

discontinuity. Contrast this with the corresponding relationships for the relaxed version in figures 3.5 and 3.6. Note that there is a loss of frequency resolution, (and corresponding increase in temporal localisation) but that this is accompanied by the avoidance of 'boundary' problems in the frequency domain; there is a smooth transition between adjacent frequency 'bins'.

The resulting tessellations of the time-frequency plane for some MFT level are shown in figures 3.7 and 3.8. Note that in the relaxed version each point on the plane falls under four analysis windows. Each coefficient is now calculated via

$$\hat{x}_{ik}(n) = \sum_{l=0}^{M-1} g'_{(\frac{i}{2}\Gamma(n)-l)}(n) x_l e^{-j\frac{2\pi}{M}\Omega(n)kl} \tag{3.4}$$

$$0 \leq i < 2^{N+1-n} \quad 0 \leq k < 2^n \quad M = 2^N \tag{3.5}$$

$$\Gamma(n) = 2^N \quad \Omega(n) = 2^{N-n} \tag{3.6}$$

In the current MFT implementation, the windows are shifted from the origin in time and frequency by one half of the sampling interval in that domain. e.g. the frequency response of the analysis

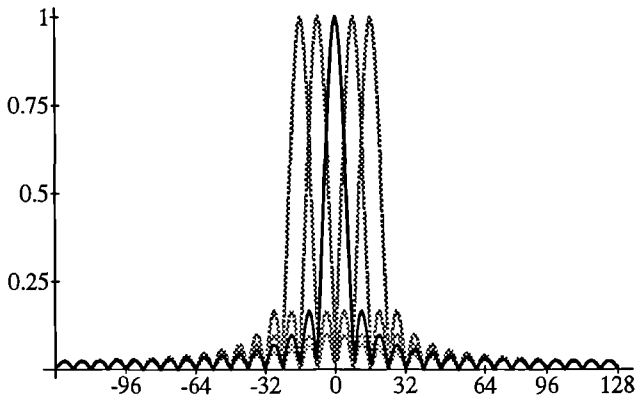

Figure 3.4: Analysis vectors (time), $\sigma = 1$

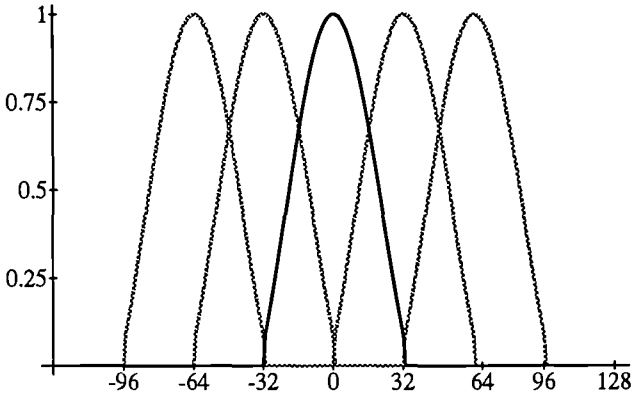

Figure 3.5: Analysis vectors (freq), $\sigma = 2$

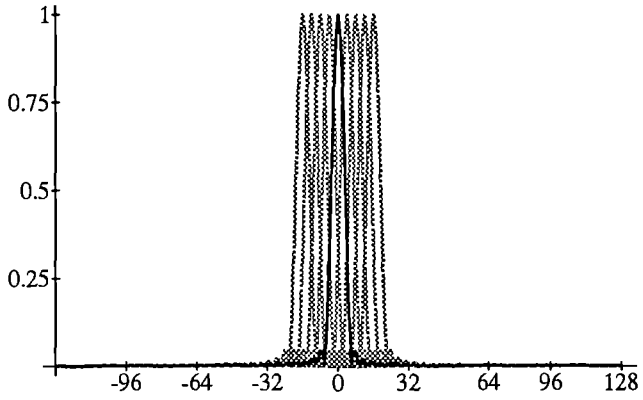

Figure 3.6: Analysis vectors (time), $\sigma = 2$

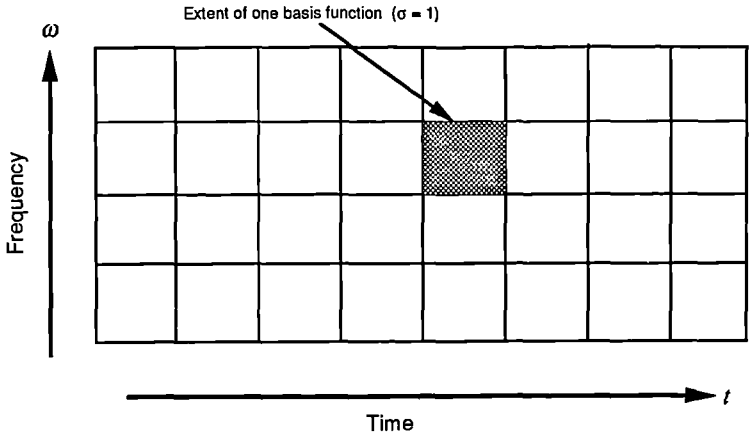

Figure 3.7: Level tessellation, $\sigma = 1$

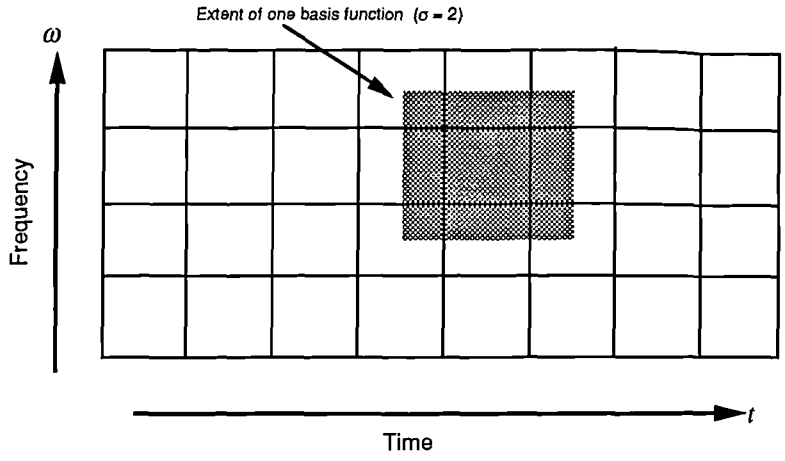

Figure 3.8: Level tessellation, $\sigma = 2$

vector for the coefficient $\hat{x}_{i0}$ has the property that

$$\hat{g}'_m = 0 \quad \text{for} \quad \frac{3\Omega(n)}{2} < m < M - \frac{\Omega(n)}{2} \tag{3.7}$$

The effect of this is to centre the coefficient 'cells' retaining the inter-level alignment between coefficient pairs described in Chapter 2 (2.48) and (2.49).

In the previous chapter it was said that a level from the non-relaxed MFT could be inverted by the application of synthesis vectors (eqn. 2.60) whose frequency response was the inverse of the synthesis vectors and that this inversion is exact. However, it can be seen that the required synthesis window (fig. 3.9) has large discontinuities in the frequency domain and so it is correspondingly poorly localised in the time domain (fig. 3.10). This lack of locality will give a poor inversion. To use the terminology of Daubechies [Dau88a], the frame is not snug. The frequency domain overlapping of analysis windows introduced into the relaxed MFT allows an alternative choice for the synthesis vector. Instead of having the inverse frequency response of the analysis vector it is now only necessary for the summation of the analysis-synthesis window products to give a 'flat' overall frequency response. Rather surprisingly, it has been found [Cal89] that a choice of

$$(\mathbf{g}')^{-1}(n) = \mathbf{g}'^{(n)} \tag{3.8}$$

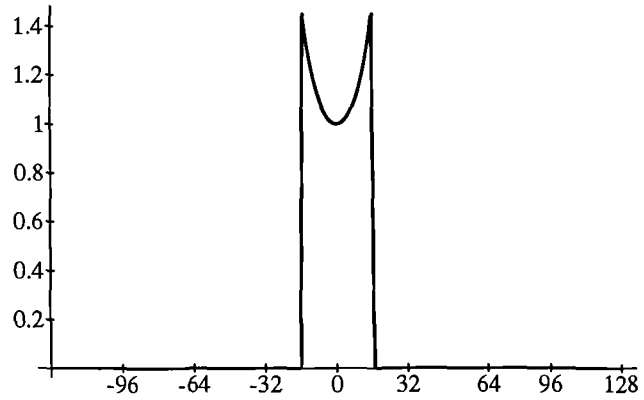

Figure 3.9: Non-relaxed MFT synthesis vector (freq)

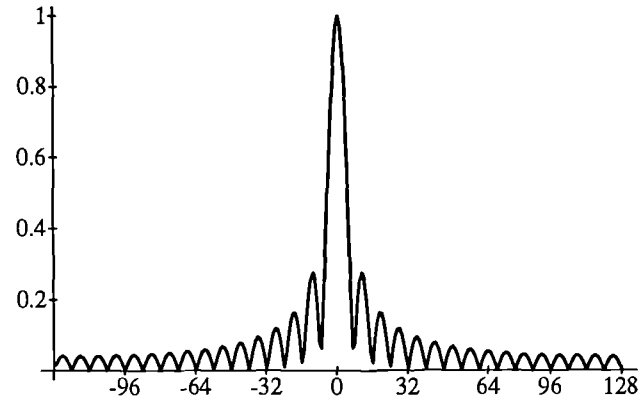

Figure 3.10: Non-relaxed MFT synthesis vector (time)

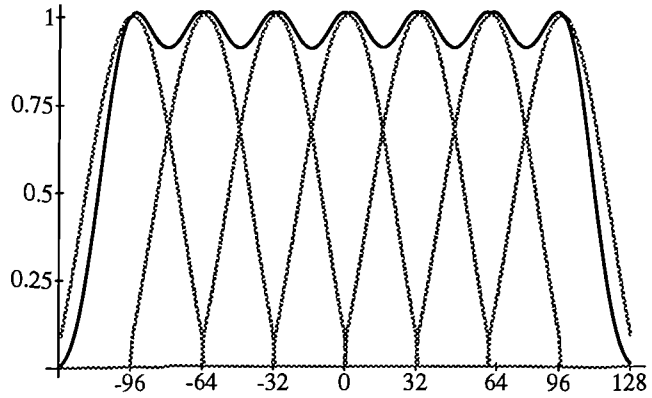

Figure 3.11: Analysis-Synthesis window products (frequency)

(i.e. reapplying the *analysis* vectors) gives a frequency response with little ripple, (fig. 3.11), and that this results in an inversion with no noticeable degradation. This result leads to the idea that the relaxed MFT analysis and synthesis windows may be approximated to in the frequency domain by the cosine based window

$$\hat{G}'_k(n) \approx \begin{cases} \cos(\frac{\pi}{2\Omega(n)}k) & |k| \leq \Omega(n) \\ 0 & \text{else} \end{cases} \tag{3.9}$$

which is more easily computed than the corresponding FPSS, though this approximation is not used in this work.

3.3 MFT Implementation

3.3.1 Analysis Vector Generation

In order to generate an MFT, it is necessary to have the appropriate set of analysis vectors available and so FPSS's of appropriate sizes must be computed. Note that these sequences can be stored and so there is no need to repeat this computation for each application of the MFT. The FPSS's are defined

by the eigenvector problem [WS87a]

$$\mathbf{B}(\Omega(n))\mathbf{T}(\Gamma(n))\mathbf{g}(n) = \lambda_0 \mathbf{g}(n) \tag{3.10}$$

(previously eqn. 2.52), where λ_0 is the largest eigenvalue of the combined operator

$$\mathbf{B}(\Gamma(n))\mathbf{T}(\Omega(n)) \tag{3.11}$$

There exist software library algorithms for the solution of eigenvector problems (e.g. [Gro89]), but for the general case they typically have a computational complexity of $O(n^3)$, which means that they are very expensive to calculate for large orders. An iterative algorithm for sequences where $\Gamma(n) = 2^r$ and $\Omega(n) = 2^s$, is suggested in [WS87a] and this has been implemented for this work. The algorithm is shown in Figure 3.12.

Stage 1 The size of the FPSS is divided by a power of two to reduce the order of the problem.

Stage 2 An FPSS of this smaller size is calculated by some library routine via Equation (3.10).

Stage 3 This sequence is expanded to the required size by successive applications of the iterative process: oversample the sequence by a factor of two and then successively apply the operations of time-truncation and bandlimiting until a convergence criterion is satisfied.

Stage 4 Finally the sequence is normalised such that its energy is unity.

The act of repetitively applying time-truncation and bandlimiting operations causes the sequence to converge such that it becomes invariant to these operations — the desired property. The technique of starting from a small FPSS and successively expanding it by a factor of two encourages convergence since there is only a small difference between the starting and final sequences at each stage.

3.3.2 Forward Transform Implementation

Implementation of the MFT itself can now be discussed. To commence, consider only the generation of one level; each coefficient can be obtained via Equation (3.4). Each row, $k = k_0$ has $2\Omega(n)$

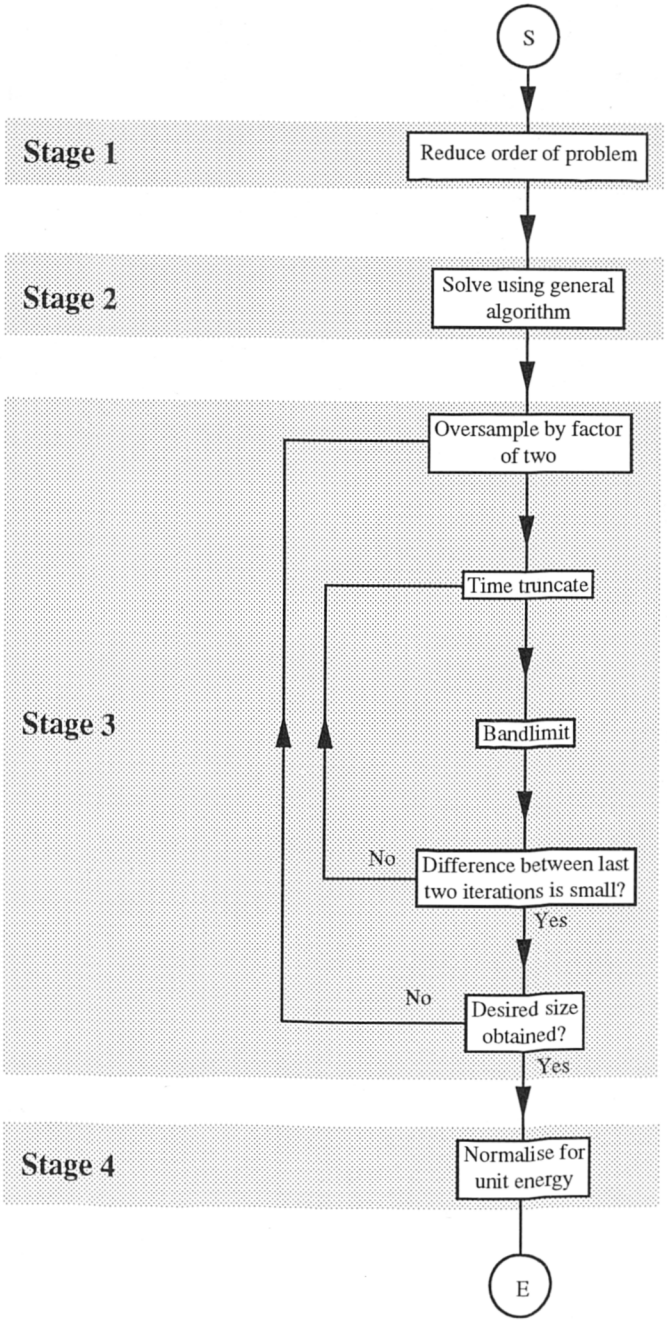

Figure 3.12: Algorithm for Efficient FPSS Generation

coefficients i.e. there is a time-domain subsampling of the M signal samples down to $2\Omega(n)$ row coefficients. The DFT of a row, with respect to i and in terms of the signal DFT $\mathbf{X}$, can be obtained directly using the subsampling and convolution properties of the DFT [Bra86]

$$\hat{x}_{\omega k_0}(n) = \frac{1}{\Gamma(n)} \sum_{r=0}^{\Gamma(n)-1} \hat{g}_{\omega+2r\Omega(n)}(n) X_{\omega+k_0\Omega(n)+2r\Omega(n)} \tag{3.12}$$

where $\hat{\mathbf{g}}$ is the DFT of the analysis vector. The summation represents the time-domain subsampling. However, recall from above that the analysis vector is truncated in the frequency domain with width $2\Omega(n)$, making the summation redundant and Equation (3.12) reduces to,

$$\hat{x}_{\omega k_0}(n) = \frac{1}{\Gamma(n)} \hat{g}_{\omega}(n) X_{\omega+k_0\Omega(n)} \tag{3.13}$$

This is simply the frequency response of the analysis vector multiplied by an appropriately shifted DFT of the signal, suggesting that the MFT can be efficiently implemented in the frequency domain, making use of the FFT [Dau88a].

The algorithm for the generation of a single MFT level is shown in Figure 3.13. To summarise:

Stage 1 Take the FFT of the entire input buffer. In this work a buffer length of 2^{16} samples is used.

Stage 2 For each frequency bin in the level, multiply the signal spectrum by the DFT of the appropriately shifted analysis vector, keeping only the samples within its support[1].

Stage 3 Take the inverse FFTs of these vectors.

Stage 4 These become the rows of the MFT level.

The generation of other levels is accomplished in a similar fashion using the appropriate analysis vector (but note that the DFT of the input buffer is common to all levels and so must only be generated once per MFT). Since the input data is real it is desirable to generate only the MFT coefficients for the positive half of the frequency axis, so reducing the storage requirements of the MFT by a factor of 2.

[1] Note that subsampling occurs implicitly here.

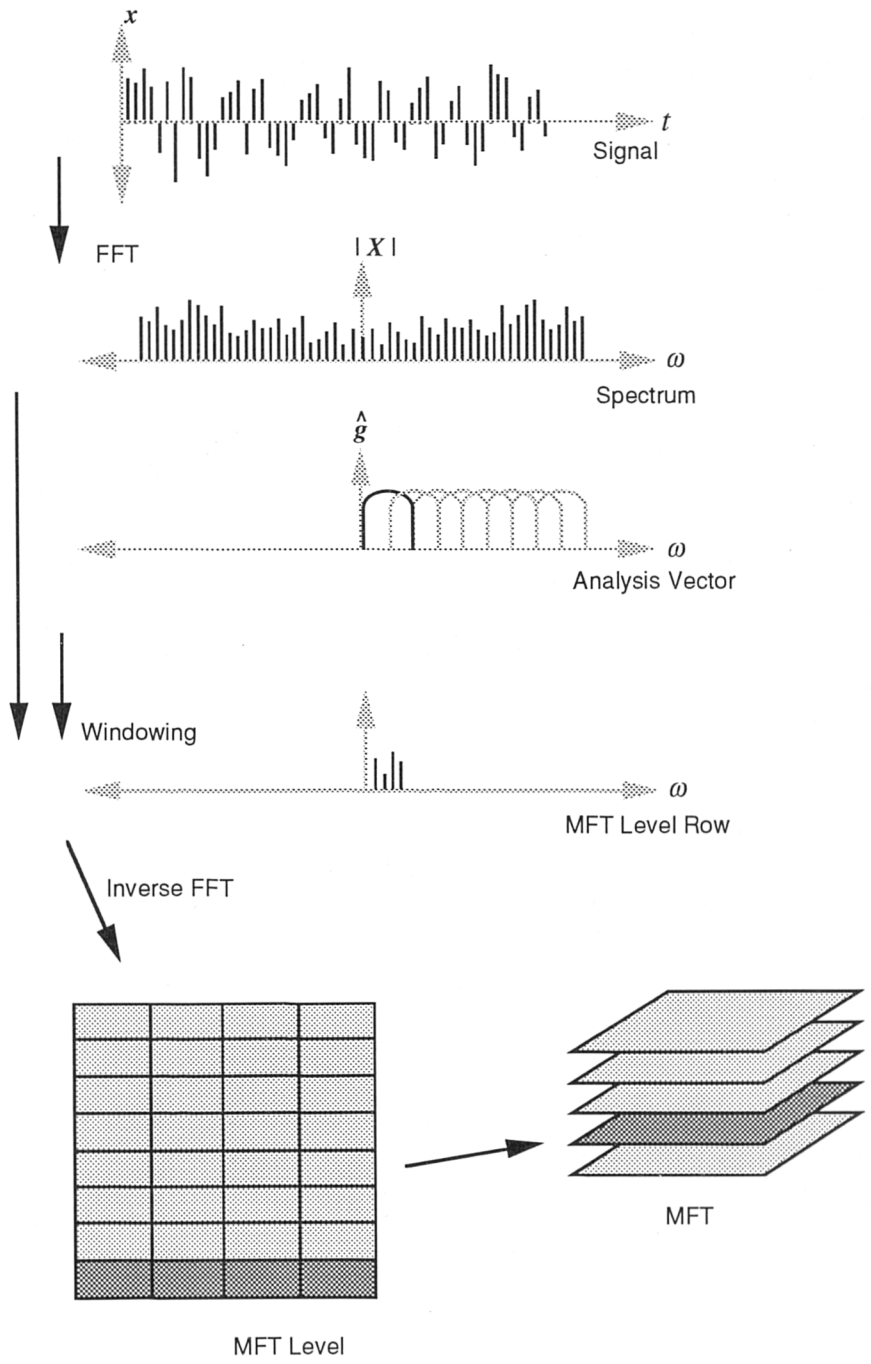

Figure 3.13: Forward Transform Implementation

In order to retain invertibility for this form it must be possible to reconstruct the whole input buffer DFT from the MFT coefficients available. To facilitate this, it is necessary to shift the input buffer DFT by one half a sample to obtain a truly Hermitian symmetric spectrum [Bra86]. The negative half of such a spectrum can then easily be obtained from the positive half. This shift is implemented in the time domain by pre-multiplying the input data by the appropriate complex exponential.

$$x'_m = x_m e^{-j\frac{\pi}{M}m} \quad 0 \leq m < M \tag{3.14}$$

Computationally, the initial DFT requires $O\left(M \log_2(M)\right)$ complex operations; for each level there are $\Gamma(n)$ vector applications and $\Gamma(n)$ order $2\Omega(n)$ inverse DFTs. Ignoring the initial DFT for a moment, the generation of an MFT level n has a computational complexity

$$O\left(\Gamma(n)(2\Omega(n) + 2\Omega(n)\log_2(2\Omega(n)))\right) \tag{3.15}$$

which reduces to

$$O\left(2M(2 + N - n)\right) \tag{3.16}$$

The requirements for the whole MFT with levels n_l to n_h are then

$$O\left(M \log_2(M) + 2M \sum_{n=n_l}^{n_h} (2 + N - n)\right) \tag{3.17}$$

This is comparable to generating the corresponding number of STFTs.

Audio signals can clearly be infinite in length, sections to be analysed typically range from 1 second upwards, and are normally sampled at rates from 8KHz up to 48KHz, giving a total number of samples $10^4 < M_{total} < 10^7$. Hardware restrictions imply that it is impractical to perform FFT calculations on very long sequences and so some algorithm is required to partition the signal into smaller sections or *blocks*. It can be seen that the lowest level required in the MFT places a lower bound on the size of this buffer by observing that its generation requires a signal DFT of at least $\Omega(n_l)$ points and that this gives an intermediate MFT with just one column. There are other issues to consider, however: in the time domain analysis vectors are not finite and adjacent vectors overlap.

While this is not normally a problem, the circular properties of the DFT imply that the first and last coefficients in a row must also be considered adjacent in time resulting in energy from the start and end of the signal appearing in the last and first few MFT columns respectively. Making sure the signal to be analysed starts and ends with a period of silence will avoid this problem in cases where the MFT is generated from a single block, but this is not possible when the signal must be transformed in many blocks. The solution used here is to use an 'overlap-save' technique by applying the cosine based window shown in Figure 3.15 to each block before transformation, transfer only the central coefficients from each intermediate MFT level to the final level and to overlap the blocks in time by 50% of their length. Note that the lower bound on block length is now increased to give an intermediate MFT with level widths of at least four coefficients. The algorithm is shown in Figure 3.14, note that this method increases the computational requirements (eqn. 3.17) of the transform by a factor of two.

3.3.3 Inverse Transform Implementation

Inversion of a single MFT level is accomplished by a process very similar to its generation. The DFT of each row of the level is taken and the analysis vector is reapplied; these vectors are then accumulated at appropriate frequencies on the signal spectrum which when inverse Fourier transformed gives the original signal.

It follows, from the discussion on the 'blocked' generation of the MFT from long signals, that some corresponding method is required for level inversion. Again the technique is similar to generation: the level is partitioned in to a series of shorter duration intermediate levels (with 50% temporal overlap), each of which is then inverted and the central samples spliced into the final signal, see figure 3.16. No windowing operation is required in this case.

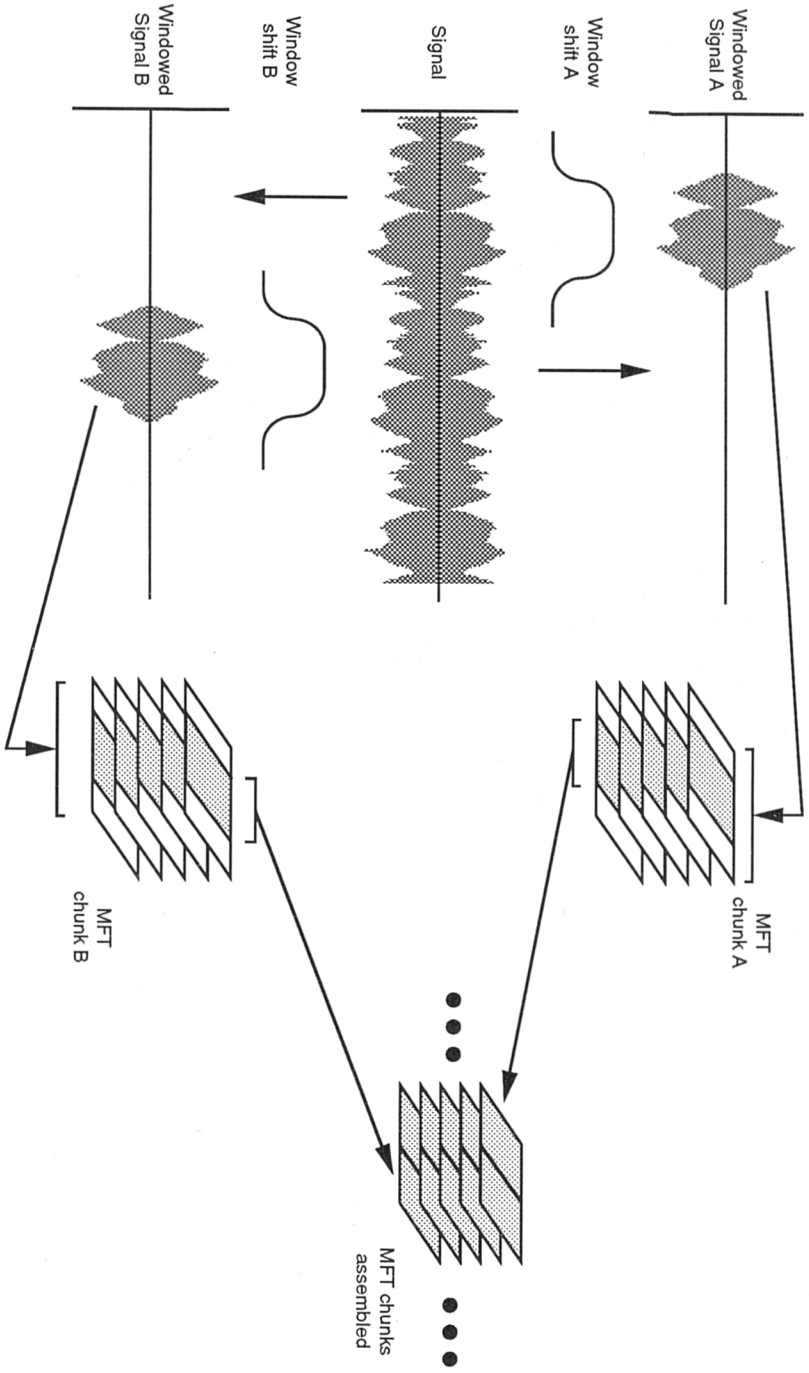

Figure 3.14: 'Blocked' Forward Transform Implementation

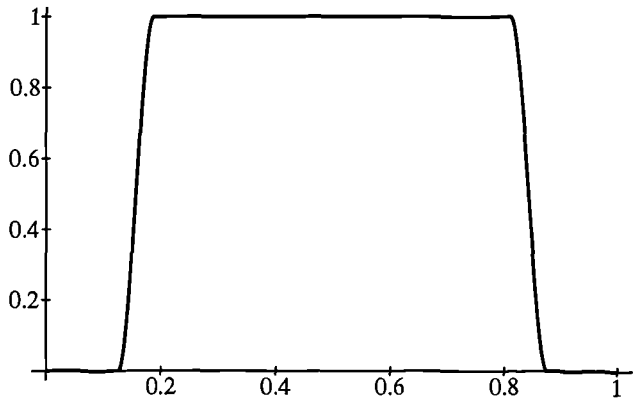

Figure 3.15: 'Blocked' Forward Transform Window

3.4 Some Examples

3.4.1 Introduction

Below are some example plots of single instrument notes transformed by the MFT. The main purpose of this section is to allow those readers who may not be familiar with time-frequency representations of musical notes to acquaint themselves with the general structure of these signals. Readers who are familiar with such plots can see that the MFT gives a familiar representation of the signals but that the multiplicity of levels delivers a more informative view of them.

3.4.2 A Simple Sinusoid

Figure 3.17 shows four MFT levels for a simple sinusoidal signal. The sinusoid has a frequency of 440 Hz. and has had a rectangular window applied to it such that it has zero magnitude before 350ms. The MFT used is, as described above, oversampled by a factor of two; notice how there are two coefficients per time frame with significant magnitude at each level. The rectangular window

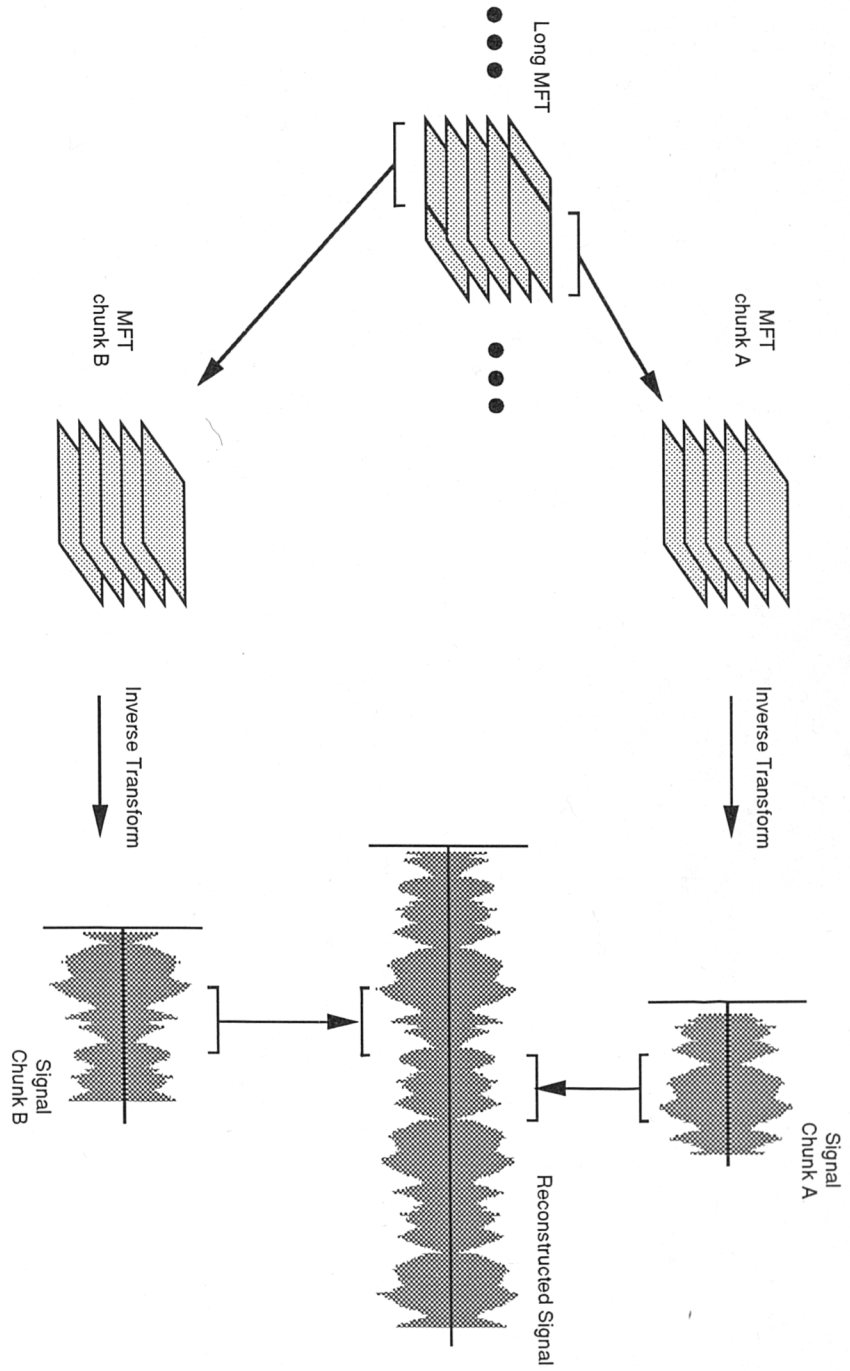

Figure 3.16: 'Blocked' Inverse Transform Implementation

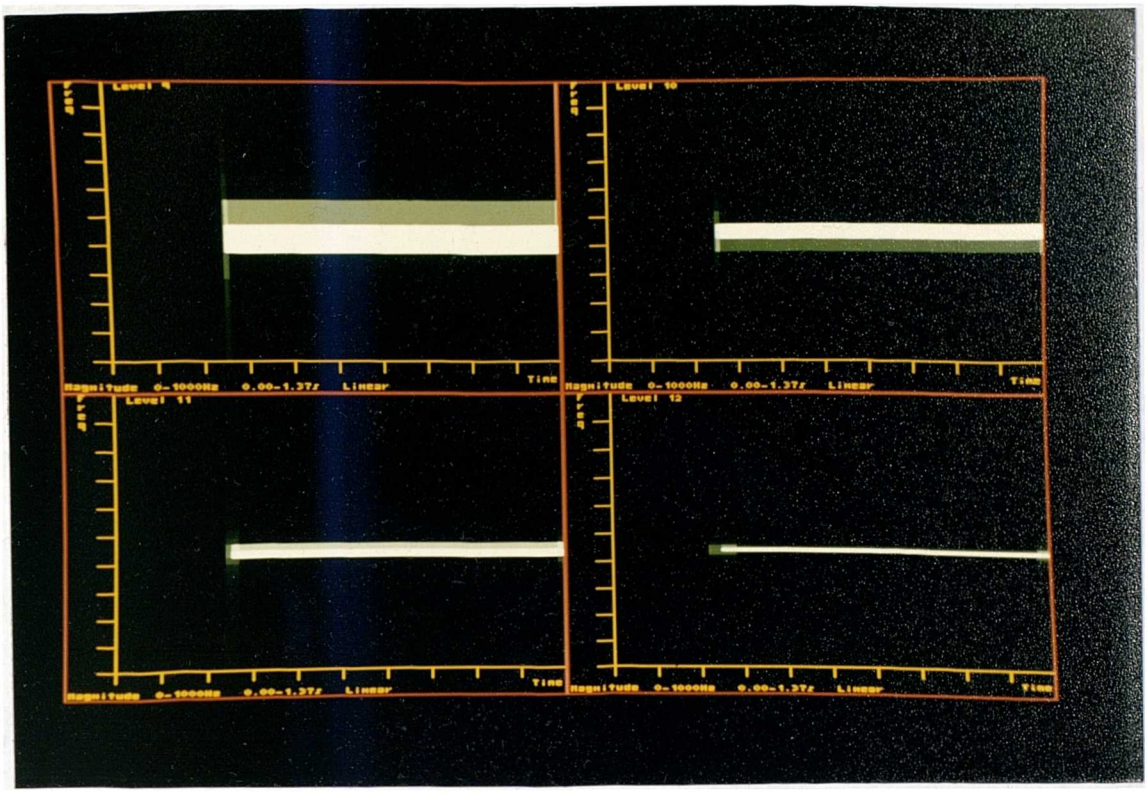

Figure 3.17: 4 MFT levels of a sinusoidal tone

results in very rapid start to the tone, (a distinct 'click' can be heard on listening). Correspondingly the energy of the signal spreads across frequency at the onset. The differences between the levels should be fairly clear: the lowest level (Level 9 - top left) has the highest time resolution enabling the time of the tone's start to be accurately observed. For each successive level (moving left to right and top to bottom), there is twice as much frequency resolution and half as much temporal resolution. The highest level gives the most accurate measure of the frequency. The higher the level, the more concentrated across frequency the sinusoid becomes: the sinusoid is well localised in frequency.

3.4.3 A complex tone

Figure 3.18 shows the same four MFT levels as before, but for a synthesised tone with five harmonics with frequencies of 110 Hz, 220 Hz, 330 Hz, 440 Hz and 550 Hz. The relative strengths of the

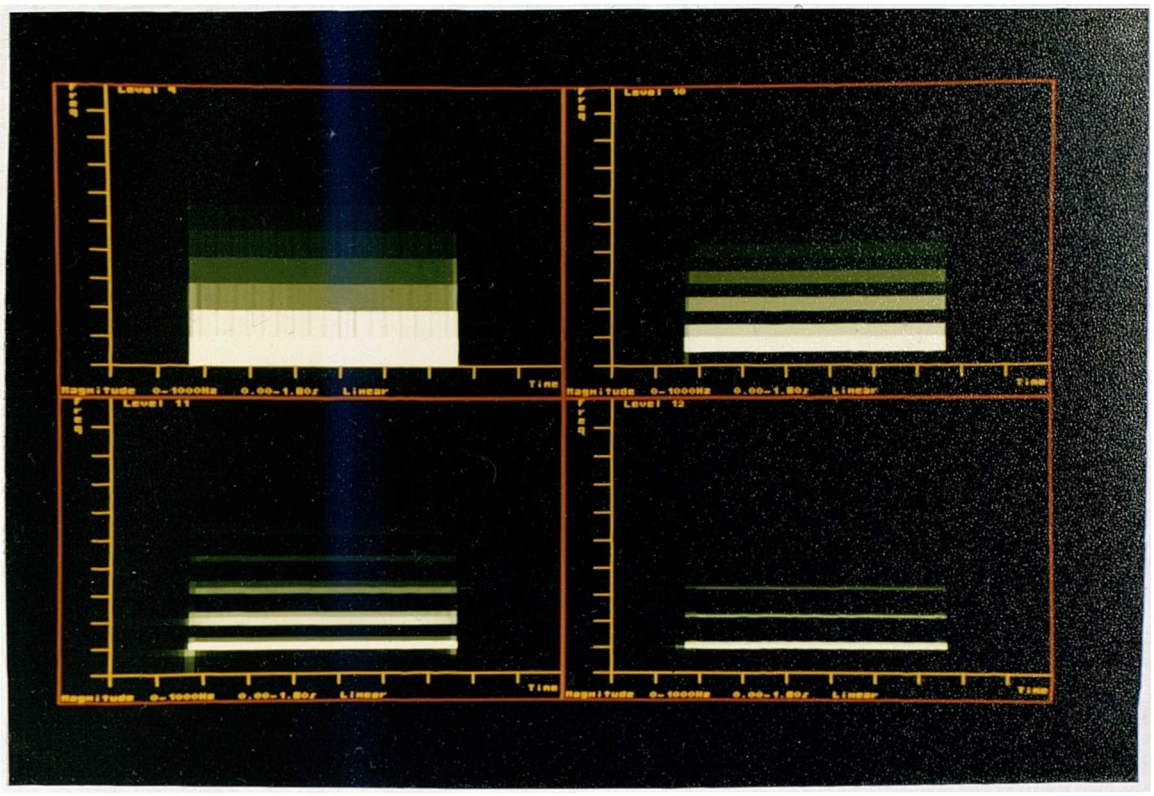

Figure 3.18: 4 MFT levels of a complex tone

harmonics are 1, 1/2, 1/3, 1/4 and 1/5.

A rectangular window was applied to the signal in the time domain before the MFT was applied such that the signal magnitude is zero before 0.3 ms and after 1.3 s; the characteristic broadband features resulting from this can clearly be seen at all levels. The most important observation to make from this signal is that at the lower two levels (top row) there is insufficient frequency resolution to separate all the harmonics from one another. In the lowest level particularly, interference can be seen as a variation in amplitude of the harmonics with time that is not present in the higher levels. Clearly, in order to obtain a meaningful representation of a signal, it is necessary to have sufficient resolution to separate its various parts.

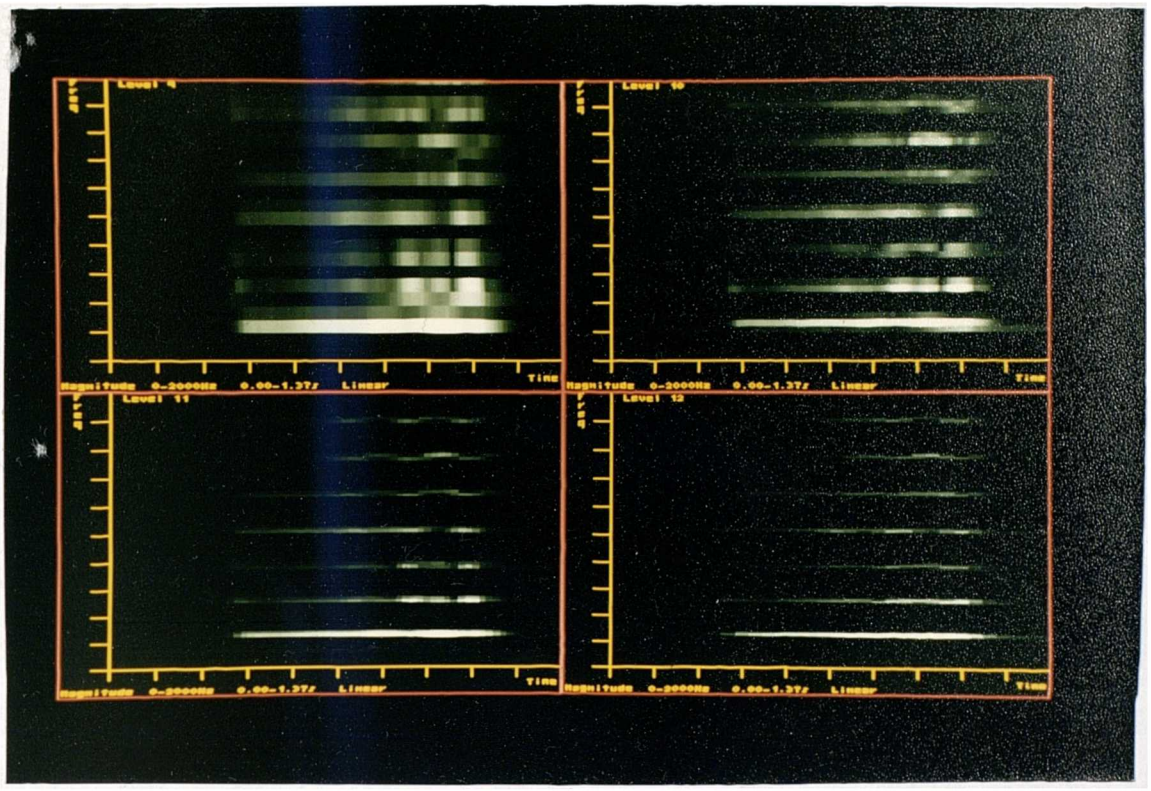

Figure 3.19: 4 MFT levels of a violin note

3.4.4 A Violin Note

Once again the same four MFT levels are shown in Figure 3.19, this time for a natural signal, a single violin note. The signal appears less 'clean' than the previous two even though it is a modern recording made in isolation. This can be attributed to the complexities of the violin as a mechanism for producing signals compared with the pure environment of a synthesiser and is typical for mechanically produced audio signals. The variation in frequency and amplitude corresponds to the presence of a small amount of vibrato in the note.

3.5 Summary

This chapter has discussed the implementation of the MFT and its application to musical signals. It has been shown that, once an appropriate set of analysis vectors has been generated, both the forward and inverse MFTs can be implemented efficiently for arbitrary length audio signals. The

computational requirements of the transform (eqn. 3.17), through use of the FFT, are similar to traditional implementations of the STFT [Cal89], though the multiple resolutions produced does imply a corresponding increase in the computational load. It is hoped that the example transforms shown above and the analysis algorithms described in the following chapters will demonstrate that the characteristics of the MFT are desirable enough to warrant the additional processing and storage required by the MFT.

The similarity of individual levels of the MFT to the STFT can only be an advantage in the area of music analysis. The STFT or phase vocoder has been widely applied by computer musicians which has resulted in a large body of experience with the transform and much knowledge of the way in which musical signals are represented by the transform. The choice of window size offered by the MFT combined with its uniform tessellation of the time-frequency plane lends itself to the application of traditional techniques. It is hoped that the methods for multiresolution analysis discussed in chapters 5 and 6 will prove useful in the adaptation of techniques developed for the STFT so that they may take advantage of the more powerful representation of signals afforded by the MFT.

Chapter 4

A Model of Note Structure

4.1 Introduction

This chapter discusses the types of features which are to be extracted and the signal properties of these features. Once the relevant parts of the signal have been identified, simple mathematical models for their behaviour in time and frequency are formed which lead to the definitions of a set of feature detection algorithms in the following chapter.

4.2 Towards A Note Detection Strategy

In order to detect and parameterise notes within musical signals, it is first necessary to observe and discuss the structure of such signals, in order that a detection strategy be formed. Aspects of individual note structure, the arrangement of notes within a polyphonic musical signal, and the presence of structures other than notes in the signal must be considered.

4.2.1 Note Structure

Notes are produced by a wide variety of instruments both natural and synthetic, each with its own distinctive timbre. An aim of this work is to detect notes from a broad range of instruments: any detection strategy having properties which are instrument specific is to be avoided. Clearly only note characteristics common across the class of source instruments may be considered as input for the detection process. Each note is defined to be a vector of parameters: the ith note is

$$\theta_i = (\theta_{i_1}, \ldots, \theta_{i_m})^T \tag{4.1}$$

The parameters, which are fully described in chapters 5 and 6, include

$$\theta_{i_1} = t_i \quad \text{the onset time} \tag{4.2}$$

$$\theta_{i_2} = f_i \quad \text{the fundamental frequency} \tag{4.3}$$

$$\theta_{i_3} = \tau_i \quad \text{the duration} \tag{4.4}$$

Using the MFT to examine notes from differing sources over a range of scales reveals the following general features

1. Notes may have one or more partials, each of which has a pseudo-sinusoidal nature.
2. The total energy of the note can be fairly arbitrarily distributed amongst its partials.
3. The onset times of the partials are grouped fairly closely.
4. The amplitude envelope of each partial can be very complex, particularly during the attack phase.
5. The frequency of the partial may deviate with time, though typically only moderately, and the deviations of partials pertaining to the same note are usually synchronised.

4.2.2 Note Juxtaposition

A musical piece can be considered to be constructed from many notes produced by several instruments arranged through time and across frequency. It is this arrangement, the score, which characterises the piece. The score may never ever exist on any media particularly when the music is composed and performed by the same artist or in the case of many forms of folk music where the music is communicated between musicians via performance. When a score does exist, it is not normally a precise definition of the signal that will be produced from it, since interpretation by the performers will 'warp' both the time and frequency parameters of the score as well as adding detail. Additionally the performance environment and recording chain will introduce effects such as reverberation and noise.

Composition normally takes place within the framework of some well defined musical system or at least within the bounds of a set of cultural traditions. These may be fairly precise, a certain pitch scale for example, or some more general structural conventions. Such rules and conventions have found their way into musical analysis systems. J. A. Moorer's system [Moo75] used a set of bandpass filters centered on the partials of the set of notes which may occur, *given* the harmony of the piece, which is estimated by a preprocessing stage using a hetrodyne filter. The typical strategy of this type of technique is to reduce the number of possible arrangements of notes at an early stage in the processing so reducing the amount of computation required by the rest of the system. Such an arrangement may be represented using a general stochastic model specified by the probability density for the ith note, $p_\theta(\theta_i|\theta_k, k < i)$, which incorporates the fact that the probability of some note occurring with a given set of parameters is affected by the parameters of the preceding notes.

For a general analysis system however, it seems undesirable to include such restrictions, particularly in the early stages of the analysis: the system will probably fail when presented with input which does not conform to the incorporated rules. With musical signals, this can be very restricting, since, particularly for more modern music, it is common practice for the composer to disregard at

least some of the traditional rules and conventions. Even if this is not the case and the music is 'well behaved', there is a danger that an erroneous decision made during preprocessing will cause the all subsequent stages of the system to malfunction.

Thus it would seem desirable to keep the feature detection strategy as independent as possible of conventions associated with any particular musical style. This system does not seek to describe the music as the score represented it but as it was performed. This description should be rich enough to allow further analysis by some subsequent processing stage which may seek to reconstruct the original score by making use of some set of stylistic conventions. Given this generality, the analysis algorithms consider the music to be a set of independent notes, i.e.

$$p_\theta(\theta_i|\theta_k, k/neqi) = p_\theta(\theta_i) \tag{4.5}$$

each of which is identically distributed

$$p_\theta(\theta_i) = p_\theta(\theta_k) \tag{4.6}$$

No assumptions are made about the relationships between each notes' parameters and so these are also considered to be both independently distributed such that

$$p_\theta(\theta_i) = \prod_{k=1}^{m} p_k(\theta_{i_k}) \tag{4.7}$$

and uniformly distributed within appropriate bounds.

Notes may thus be laid out at any position on the time frequency plane, which leads to the following observations:

1. When two or more notes of differing pitches are active at any time their partials become meshed across frequency; neighbouring partials may not be attributable to the same note.

2. Partials, or sections of partials, belonging to distinct notes may be coincident on the time frequency plane.

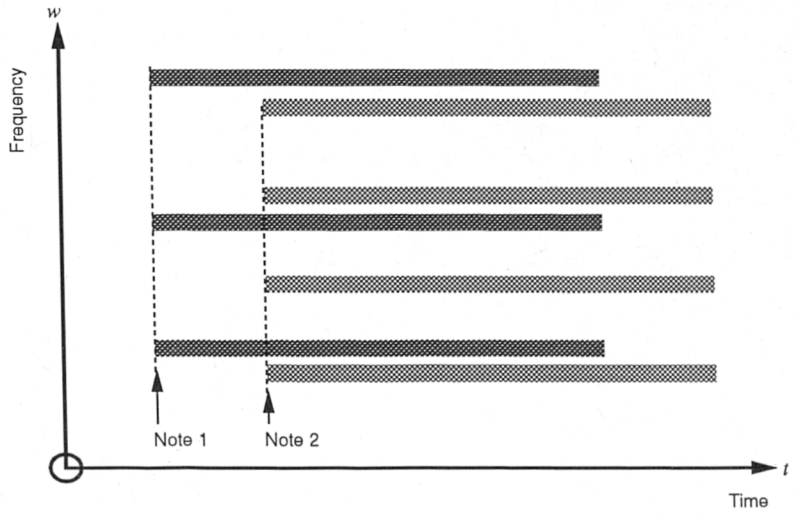

Figure 4.1: Partial Meshing

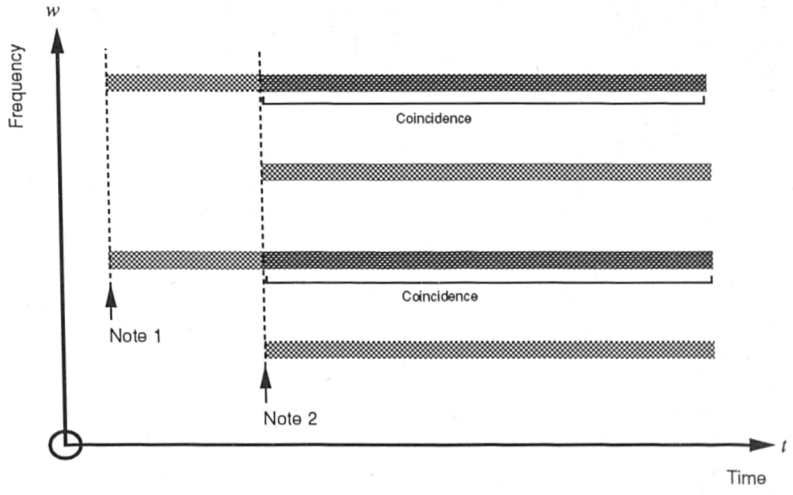

Figure 4.2: Partial Coincidence

The meshing of partials (fig 4.1) implies that it is impossible to find a rectangular region of the time frequency plane which contains all the energy from a given note and energy from no other notes. Clearly a simple algorithm for isolating notes will be insufficient in such cases. The second problem, coincidence (fig 4.2), is rather more serious. The implication here is that it is impossible to find any *set* of time-frequency regions which between them contain all the energy from a given note and energy from no other notes. Note isolation purely by means of time-frequency localisation will be impossible in such cases and this may lead to ambiguities in the detection process. In fact this situation is common, particularly when the music is bound by some melodic framework, which typically constrains the choice of pitches for the notes and defines simple frequency ratios between those pitches. The most extreme, though still common, case of this occurs when two instruments play simultaneously one octave apart: all of the partials of the higher note will be coincident with the even numbered harmonics of the lower note.

4.2.3 Other Signal Features

A musical signal may contain features other than the notes discussed above. These features may be broadly classified as follows:

1. Inharmonic sounds produced by percussive instruments such as cymbals or snare drum which are broad-band.

2. Highly complex sounds such as those used in the Musique Concrète school of composition e.g. engine noise, breaking waves and other complex textures.

3. Artifacts introduced by the recording or transmission chain through which the signal has passed.

Recognition of these structures is outside the scope of this work and so the source material is restricted to pieces not containing percussive or highly complex sources. Otherwise there are few

restrictions, poorly recorded or badly distorted signals are avoided, material is taken from Compact Discs of live and studio recorded work. Room reverberation and other such multi-path ambient effects (whether natural or synthetic) are not explicitly avoided as was the case with the resynthesised material used in [Wat86].

4.3 A Feature Hierarchy

The preceding discussion implies that the total energy of a note may be distributed over a reasonably large area of the time frequency plane, and the total energy within that area may be attributable to several notes and other features. The strategy used in this work relies on identifying a set of isolated signal features on the time-frequency plane by finding analysis scales (the multiple resolutions of the MFT) which best localise these features, allowing them to be parameterised independently. A note is clearly very poorly localised: to consider it a *single* feature could almost certainly make impossible the task of choosing an appropriate scale from which its parameters can be estimated. Choice of appropriate scale is discussed at length in Chapter 6, first it is necessary to identify a set of *localised* features which together correspond to our perception of a note; the obvious method for this is to decompose the note structure.

The primary components of a note are its partials: a set of time-varying pseudo-sinusoidal waveforms having frequencies in a nearly harmonic sequence. This harmonic structure is represented by a set of parameter vectors θ_{ij} for each note θ_i. Again these parameters are fully described in chapters 5 and 6 but include

$$\theta_{ij_1} = t_{ij} \quad \text{the onset time} \tag{4.8}$$

$$\theta_{ij_2} = f_{ij} \quad \text{the onset frequency} \tag{4.9}$$

$$\theta_{ij_3} = \tau_{ij} \quad \text{the duration} \tag{4.10}$$

While being far more localised in frequency than a note, each may be long in duration and highly

time variant. Previous workers e.g. [Gre75] have shown that piecewise linear approximations to the amplitude and frequency trajectories of partials result in a compact signal description which retains many of the perceptually significant details of the signal. However, it has been noted e.g. [Ser89], that for a general system, it is not sufficient to have fixed length linear segments to give compact descriptions of steady-state sections while retaining the ability to represent accurately the more time-variant sections such as the attack phase and vibrato. In this work, partials are broken down into a set of *partial segments* each of which maps onto a small number of coefficients from a single MFT level. Thus the duration of the segments varies according to which MFT level they are associated with. With this scheme it is possible to seek compact partial descriptions, such as would be useful for data compression, but the system presented here selects segment durations in order to extract a highly accurate description of the partial's time-frequency evolution.

The importance of the onset of a note has been reported by several workers e.g. [GM77]: these sections of a note hold important cues for the perception of timbre and source identification. While these aspects of signal analysis do not fall within the scope of this analysis system, it is clear that the onset should be described as accurately as possible in order that this important information is retained for any subsequent analysis stages. In contrast to the subsequent sections of a note, the attack phase may be highly time-variant, particularly for the fast attacks produced by striking or plucking a string. A consequence of uncertainty is that, in contrast to a partial segment, this energy will be relatively poorly localised in frequency. For this reason and because of its great perceptual importance, the signal model of a partial used here incorporates additional information about the onset energy of a partial. Experimentation has shown, however, that for many instruments the quantity of this energy is small compared with the total note energy, making detection difficult, and that, for relatively slow onsets, it may be negligible.

The proposed feature hierarchy is shown in Figure 4.3. The score consists of a set of *notes* each of which is decomposed into a set of *partials*. Each partial maps onto some set of MFT coefficients,

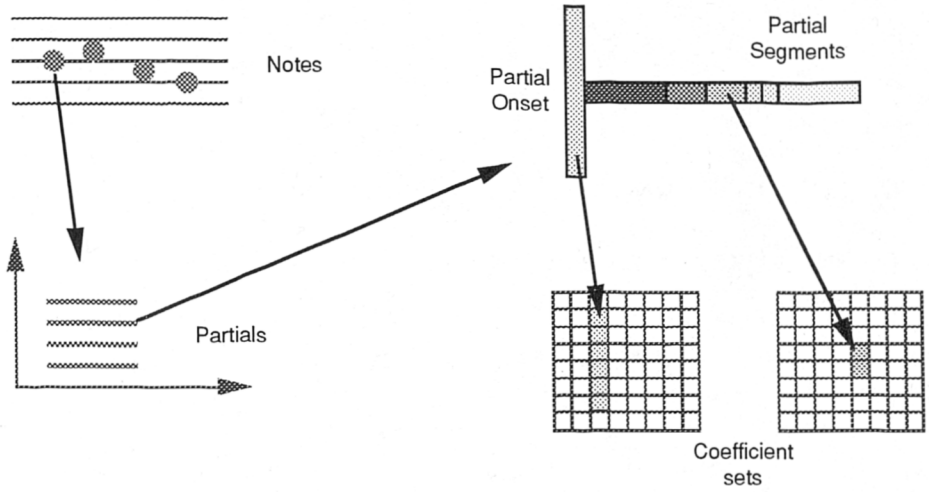

Figure 4.3: The Feature Hierarchy

which may not be distinct due to partial overlapping, via a sequence of *partial segments* and an optional *partial onset*. Since the partial components map directly on to MFT coefficients, they are thus the features which are to be extracted from the MFT. It should be noted that these features are far better suited to this task than those in the higher levels of the hierarchy as they are reasonably well localised on the time-frequency plane.

4.4 A Signal Model

In order to design suitable feature detectors it is prudent first to form a mathematical model of those features. The purpose of this model is not to provide a comprehensive definition of the source data since this would be far too complex to be of practical value. The signal model should be a relatively simple, though not unrealistic, symbolic representation of the class of target feature. In practice the actual algorithms used for the feature detectors may be somewhat more complex than those required by the signal model. The purpose of the model is to simplify the general problem in order that an initial set of definitions for the detector algorithms may be deduced.

4.4.1 Notes

The signal for a note, $N(t)$, is simply the sum of its partials

$$N(t) = \sum_{k=1}^{H} p_k(t) \tag{4.11}$$

where $p_k(t)$ is the kth partial of $N(t)$. A note can be considered in some sense *ideal* when the frequencies of its partials, $\omega_k(t)$, form a harmonic series

$$\omega_k(t) = k\omega_N(t) \tag{4.12}$$

where $\omega_N(t)$ is the fundamental frequency of the note at time t. In terms of the note and partial parameter sets θ_i and θ_{ik} this relationship gives

$$f_{ik} = kf_i \tag{4.13}$$

In addition such an ideal note will have the onset times of its partials perfectly synchronised

$$t_{ik} = t_i \tag{4.14}$$

and all partials will have similar durations

$$\tau_{ik} = \tau_i \tag{4.15}$$

In practice, however, due both to noise in the signal and the complexities of many instruments, such simple relationships are rarely observed and a more relaxed model is required. Variation in the estimates of these parameters can be modelled by using Markov processes [Pap84]. The onset times of a note's partials are then related via

$$t_{ik} = t_{i(k-1)} + \epsilon_t \quad k > 1 \tag{4.16}$$

where ϵ_t is a random variable with zero mean and variance such that $|\epsilon_t| \ll \tau_{ik}$.

Similarly the frequencies of the partials are modelled using

$$f_{ik} = \begin{cases} \frac{k}{k-1} f_{i(k-1)} + \epsilon_k & k > 1 \\ f_i + \epsilon_k & k = 1 \end{cases} \tag{4.17}$$

where ϵ_k are random variables with zero mean and variance such that $|\epsilon_k| \ll f_{ik}$. While this form is general enough to represent a wide range of instruments, it should be noted that cases have been reported for specific instruments (e.g. the piano [Bla65]) where a systematic bias in the values of ϵ_k can be detected.

Finding a model which is reasonably instrument independent for the amplitudes of the partials has proved not to be so straightforward. Great variation in the partial amplitudes can be observed both between instruments and even for different note instances produced by the same instrument. Thus without introducing specific instrument models into the system, little can be deduced from the relative amplitudes of the partials. However a form of Markov model has been used to describe the normalised onset magnitudes used in the onset detector algorithm (see Section 6.3.5).

4.4.2 Partials

The signal for each partial is modelled as the sum of two functions

$$p(t) = v'(t) + o(t) \tag{4.18}$$

where $o(t)$ is an optional short-time 'onset function' while $v'(t)$ represents the 'steady-state' portion of the partial and can be modelled in the time domain as a slowly-varying sinusoidal (pseudo-sinusoidal) function

$$v'(t) = a_{v'}(t)\cos(t\omega_{v'}(t) + \phi_{v'}) \tag{4.19}$$

where $a_{v'}(t)$ is the amplitude envelope of v' and $\omega_{v'}(t)$ is its frequency at time t. The phase offset $\phi_{v'}$ is constant for each v'. Such sinusoidal forms have been used on many occasions for the modelling of speech e.g. [Por81, MQ86], musical partials [GG78, Ser89, and others] and have also been used successfully for non-harmonic sounds [SS86]. The envelope functions for amplitude $a_v(t)$ and pitch $\omega_v(t)$ are defined to be slowly time-varying functions, and so they have the majority of their spectral energy far removed from the frequency of the carrier they modify.

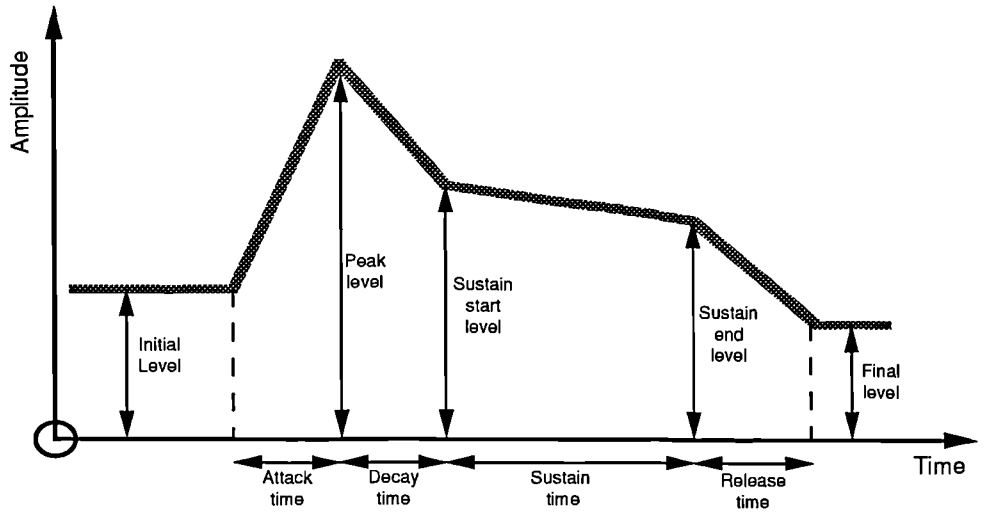

Figure 4.4: Typical Synthesiser Envelope Model

It is common practice to work with complex exponential signals instead of sinusoids in order to simplify the modelling process. Therefore Equation (4.19) is rewritten as

$$v'(t) = \Re(v(t)) \tag{4.20}$$

$$v(t) = a_v(t)e^{j\omega_v(t)t+\phi_v} \tag{4.21}$$

The complex form is used from now on; its relationship to the original form can be easily seen.

The frequency of the partial is expected to remain fairly constant with time, deviating only slightly from this mean value. Vibrato is the most common form for this modulation, it has been reported [Sma37] for a typical case on the violin as being approximately sinusoidal with frequency 6-7Hz and deviation 5% of the carrier frequency.

The amplitude envelope is far more significant and is often referred to as just the envelope of the partial. Generally speaking, its shape may be fairly arbitrary, typically rising quickly to its maximum then decaying to some lower value at which it remains before finally being damped to zero. The model shown in Figure 4.4 has been used by many synthesiser manufacturers for several years and has been found general enough to simulate a wide range of instruments. Most modern synthesisers allow all of the envelope's breakpoints to be adjusted in height and time, though for

many common instruments sounds, the settings follow the shape shown in Figure 4.4. It can be seen that $v(t)$ alone is insufficient to represent this envelope model, the rapidly increasing attack section cannot be accounted for by the slowly-varying nature of $a_v(t)$. The 'onset function', $o(t)$, models this transient element at the start of a partial which cannot be accounted for by $v(t)$. Clearly, for the partial to be represented accurately over its whole length, two distinct model elements are required, and this reflected in Equation (4.18). Recently there has been a trend to use short recordings of the attacks of real instruments to supplement the attack and decay phases of these piecewise linear synthesiser envelopes. The great increase in realism which results from this addition suggests that such a simple model is insufficient to represent the attack complexities of 'natural' instruments, whereas the latter portions of the note are adequately synthesised. Experimentation with this form of synthesis soon reveals that the choice of attack used affects the timbre of each note much more than changes to the synthesised part, indicating that the majority of instrument dependent perceptual cues are contained within the attack. This phenomenon was described by [GM77], who reported the significant effects of removing these parts from recorded instrument sounds. The proposed partial model does not account for all such energy: the definition of $o(t)$ (below) represents only the simple transient energy present at the partial onset. Any more complex features will be left in the residual energy after that accounted for by the partial model has been removed. The use of such residual energy was demonstrated in [Ser89] where it was set aside for use by a resynthesis system. However, these parts of the signal are not analysed further in this work.

The slowly time-varying nature of $v(t)$ leads to the observation than within the region of some time t_x the magnitude and frequency of the partial will remain fairly constant

$$a_v(t_x + \tau) \approx a_v(t_x) \tag{4.22}$$

$$\omega_v(t_x + \tau) \approx \omega_v(t_x) \tag{4.23}$$

$$|\tau| < T_v(t_x) \tag{4.24}$$

where $T_v(t)$ is some scale related constant for the partial at time t. Consequently Equation (4.21)

may be locally approximated by

$$v(t_x + \tau) \approx a_v(t_x)e^{j(t_x+\tau)\omega_v(t_x)+\phi_v} \qquad |\tau| < T_v(t_x) \tag{4.25}$$

The partial $v(t)$ is approximately sinusoidal in the region of t_x.

Consider now the STFT representation $V_2(t,\omega)$ of $v(t)$ about time t_x, using an analysis window $h(t)$

$$V_2(t_x + \tau, \omega) = \int_{-\infty}^{\infty} h(t_x + \tau - \mathcal{T})v(\mathcal{T})e^{-j\omega\mathcal{T}}\, d\mathcal{T} \tag{4.26}$$

Using the steady-state approximation in Equation (4.25) while constraining the duration of the analysis window to be shorter than the scale duration constant $T_v(t_x)$ allows $V_2(t_x + \tau, \omega)$ to be considered narrow-band in the region of $\omega_v(t_x)$, giving

$$V_2(t_x + \tau, \omega) = 2\pi a_v(t_x)e^{-j((t_x+\tau)\omega_v(t_x)+\phi_v)}H(\omega_v(t_x) - \omega) \tag{4.27}$$

That is the Fourier transform, $H(w)$, of $h(t)$ shifted to $\omega_v(t_x)$, weighted by the amplitude envelope at time t_x and multiplied by a complex exponential. The phase of $V_2(t,\omega)$ in the region of t_x and $\omega_v(t_x)$ is

$$\begin{aligned} \varphi_v(t_x + \tau) &= \arg\left(V_2(t_x + \tau, \omega_v(t_x))\right) \\ &= -(t_x + \tau)\omega_v(t_x) + \phi_v \end{aligned} \tag{4.28}$$

Note that, locally, $\varphi_v(t)$ is a linear function of t; it has be shown [Por81] that its derivative

$$\tilde{\varphi}_v(t) = \omega_v(t_x) \tag{4.29}$$

can be considered to be the *instantaneous frequency* of v at time t_x.

The onset function $o(t)$ will be highly instrument dependent and will in general have increasing total energy as the onset time of the partial decreases. The simplest form for the onset of a partial is a step function giving a 'rectangular' envelope. This form, however, is only produced by the most rudimentary electronic instruments. In general the form of $o(t)$ is unknown and may be highly

complex, precluding symbolic analysis, though some general observations can be made. Assuming that $o(t)$ is a short-time function with centroid at $t = \eta_o$, then its Fourier transform $O(\omega)$ will have the property that [Pap77]

$$\arg\left(O(\omega)\right) = \omega\eta_o + \phi_o \tag{4.30}$$

where ϕ_o is some phase constant. In other words the position of $o(t)$ is directly proportional to the phase variation of its spectrum.

These signal models are now considered in the sampled domain by regularly sampling the STFT of the signal with sampling intervals of Γ and Ω in time and frequency respectively. Applying this sampling to the signal model of a partial (eqn. 4.27) across time in the region of t_x at the frequency of the partial, $\omega_v(t_x)$, gives a set of complex samples $\hat{s}_{ik_v}$, $1 \leq i \leq I$. The linear form of Equation (4.28) then implies that the value of the sample with time index i depends only upon the value of the previous sample $\hat{s}_{(i-1)k_v}$. Once again, a suitable model for this behaviour is a first-order autoregressive or Markov process [Pap84]

$$\hat{s}_{ik_v} = \begin{cases} \alpha\hat{s}_{(i-1)k_v} + \beta\epsilon_i & i > 1 \\ \epsilon_1 & i = 1 \end{cases} \tag{4.31}$$

where the regression coefficient

$$\alpha = |\alpha| \exp\left(-j\omega_v(t_0)\Gamma\right) \tag{4.32}$$

and

$$0 \leq |\alpha| < 1 \quad |\alpha|^2 = 1 - \beta^2 \tag{4.33}$$

The complex values ϵ_i, representing noise in the sample estimates, are normally distributed with zero mean

$$E\left[\epsilon_i\right] = 0 \tag{4.34}$$

and unit variance

$$E\left[\epsilon_i \epsilon_l^*\right] = \delta(i - l) \tag{4.35}$$

Given this, it has been shown [Pap84] that

$$E\left[\hat{s}_{ik_v}\hat{s}^*_{lk_v}\right] = \alpha^{i-l} \quad i \geq l \tag{4.36}$$

Thus, if a partial is present, this relationship may be used as a basis from which estimates of α and hence $\omega_v(t_x)$, the partial frequency, may be made.

A similar form can be used to model the short-time onset signal $o(t)$. This is suggested by the linear form of Equation (4.30). Sampling the MFT of $o(t)$ across frequency in the temporal neighbourhood of its centroid η_o, with sampling intervals as above, will give a set of complex samples $\hat{s}_{i_\eta k}$, $1 \leq k \leq K$, where

$$\hat{s}_{i_\eta k} = \begin{cases} \alpha\hat{s}_{i_\eta(k-1)} + \beta\epsilon_k & k > 1 \\ \epsilon_1 & k = 1 \end{cases} \tag{4.37}$$

where the regression coefficient is now

$$\alpha = |\alpha| \exp(-j\eta\Omega) \tag{4.38}$$

and the other parameters are as described above.

These processes assume a phase coherent model. While not strictly necessary, this approach has been adopted to investigate the use of such information and to confirm that, since it is new, the MFT presents this information in a suitable manner. The limitations of this approach is that phase coherence may be lost when several features, such as partials or onsets, are coincident. This is discussed further in chapters 7 and 8.

4.5 Summary

This Chapter has proposed a feature hierarchy which relates the coefficients of the MFT through a set of localised features to the notes and ultimately the score of a piece of music. The features which map directly onto the MFT coefficients have the important property of being localised in both

time and frequency allowing them to be isolated from one another by the use of efficient small neighbourhood detectors.

Having established a feature hierarchy, models for each of the elements were defined allowing various parameter and coefficient relationships to be described in terms of first-order autoregressive processes.

The following two chapters define a set of feature detection algorithms based on these models and describe how these are integrated into a multiresolution framework for the analysis of polyphonic music signals.

Chapter 5

Feature Detection

5.1 Detection Overview

This is the first of two chapters describing the analysis algorithms. In this chapter, algorithms are developed to implement detectors which attempt to recognise the features described in the preceding chapter. An overview of the analysis procedure is shown in Figure 5.1.

The detectors are based upon the signal models developed for the corresponding features with the addition that the data on which they operate is modelled as the sum of the signal and independent white noise w with variance σ^2

$$\mathbf{x} = \mathbf{s} + \mathbf{w} \tag{5.1}$$

The raw data is first transformed using the MFT to give sets of level coefficients

$$\hat{\mathbf{x}}(n) = \mathbf{F}(n)\mathbf{x} \tag{5.2}$$

Subsequent processing falls into two main categories: that performed on data from each MFT level independently and that which combines estimates made across a range of scales. This chapter deals with the per level processing, the next chapter deals with the scale space based processing.

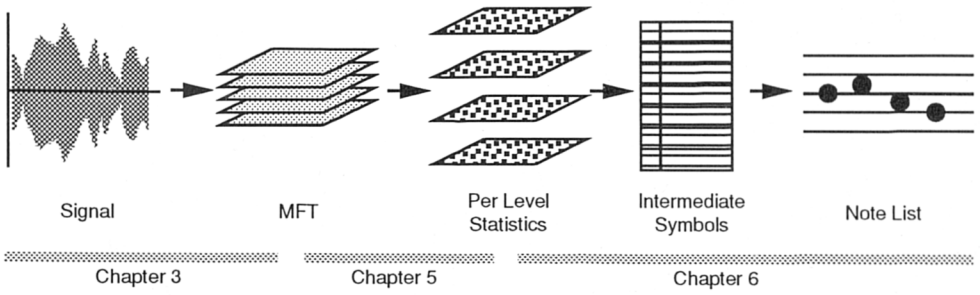

Figure 5.1: Analysis Overview

5.2 Feature Detection

A model for the slowly-varying element of a partial was described in the previous chapter (eqn. 4.19) and it was shown that, over short periods, it had constant frequency and magnitude and so may be represented in that region by a single complex exponential. A discrete form of this signal approximation may be written, simply, as the sequence

$$\mathbf{s} = \{s_m : 0 \leq m < M\} \tag{5.3}$$

where

$$s_m = a_s e^{-j\frac{2\pi}{M}\omega_s m} \tag{5.4}$$

and a_s is the locally constant partial amplitude. The angular frequency ω_s is related to the frequency of the partial f_s Hz. via $\omega_s = f_s/f_S$ where f_S is the sampling frequency (48 kHz.)

An oversampled version of the MFT is used in this work, the inequality in Equation (2.45) becomes

$$N_i(n)N_k(n) = 2M \tag{5.5}$$

Chapter 3 gave a definition of this oversampled MFT (eqn. 3.4) and described several advantages of using this form. The coefficients of this MFT for the input sequence s are given by

$$\hat{s}_{ik}(n) = \sum_{l=0}^{M-1} g'_{(l-\Gamma(n)/2)(i+\frac{1}{2})}(n) s_l e^{-j\frac{2\pi}{M}\Omega(n)(k+\frac{1}{2})l} \tag{5.6}$$

where all indices are calculated modulo M. For the steady-state partial sequence, this gives coefficients

$$\hat{s}_{ik}(n) = a_s \sum_{l=0}^{M-1} g'_l(n) e^{-j\frac{2\pi}{M}((k+\frac{1}{2})\Omega(n)-\omega_s)(l+(i+\frac{1}{2})\frac{\Gamma(n)}{2})} \tag{5.7}$$

$$= a_s \hat{g}'_{(k+\frac{1}{2})\Omega(n)-\omega_s}(n) e^{-j(\frac{2\pi}{M}((k+\frac{1}{2})\Omega(n)-\omega_s)(i+\frac{1}{2})\frac{\Gamma(n)}{2}} \tag{5.8}$$

The coefficient magnitudes are given by

$$|\hat{s}_{ik}(n)| = a_s \hat{g}'_{(k+\frac{1}{2})\Omega(n)-\omega_s}(n) \tag{5.9}$$

which is independent of the time frame. The analysis windows are exactly bandlimited in the frequency domain with bandwidth of $\frac{4\pi}{M}\Omega(n)$, so that

$$\hat{g}'_m(n) = 0 \quad \text{for} \quad |m| > \Omega(n) \tag{5.10}$$

The spacing between adjacent frequency bins is only $\Omega(n)$ giving a 50% overlap with the result that there are always two non-zero coefficients associated with the partial. This was illustrated in Figure 3.5 which showed the frequency arrangement of the analysis windows in each time frame. The two non-zero coefficients are $\hat{s}_{ik_1}$, referred to as the *primary bin*, and one of its neighbours, the *secondary bin*, $\hat{s}_{ik_2} = \hat{s}_{i(k_1 \pm 1)}$. The frequency index k_1 of the primary bin satisfies

$$0 \leq \frac{\omega_s}{\Omega(n)} - k_1 < 1 \tag{5.11}$$

The frequency response of the analysis window results in

$$|\hat{s}_{ik_1}| \geq |\hat{s}_{ik_2}| \tag{5.12}$$

The approximation of the analysis window frequency response by a cosine based function in Chapter 3 (eqn. 3.9) allows the magnitude of the partial to be readily obtained via

$$a_s \approx K_n \sqrt{|\hat{s}_{ik_1}|^2 + |\hat{s}_{ik_2}|^2} \tag{5.13}$$

where K_n is a constant for each MFT level.

The phase of the coefficients in Equation (5.8) is given by

$$\arg(\hat{s}_{ik_1}(n)) = -(\frac{2\pi}{M}((k_1 + 0.5)\Omega(n) - \omega_s)(i + 0.5)\frac{\Gamma(n)}{2} \tag{5.14}$$

As was shown in Chapter 4, a measure of the instantaneous frequency of a partial can be determined by the derivative of its phase with time. Taking the forward phase difference (eqn. 2.64) of successive time frames at the partial's primary bin gives

$$\begin{aligned} \nu_{ik_1}(n) &= \arg(\hat{s}_{(i+1)k_1}(n)) - \arg(\hat{s}_{ik_1}(n)) & (5.15)\\ &= -\frac{2\pi}{M}((k_1 + 0.5)\Omega(n) - \omega_s)\frac{\Gamma(n)}{2} & (5.16)\\ &= -\frac{\pi}{\Omega(n)}((k_1 + 0.5)\Omega(n) - \omega_s) & (5.17) \end{aligned}$$

When the partial's frequency lies at the centre of its primary bin $\omega_s = (k_1 + 0.5)\Omega(n)$ then $\nu_{ik_1}(n) = 0$. The maximum frequency deviation of the partial from this point is $\frac{2\pi}{M}\Omega(n)/2$, any greater deviation will (eqn. 5.11) cause another frequency bin to be chosen as the primary bin. Consequently

$$|\nu_{ik_1}| < \frac{\pi}{2} \tag{5.18}$$

allowing the partial frequency to be determined unambiguously via

$$f_s = (\frac{\nu_{ik_1}}{\pi} + k_1 + 0.5)\frac{\Omega(n)}{M}f_S \tag{5.19}$$

In the case that the frequency of the partial remains constant over the duration of several time frames then

$$\nu_{ik_1} = \nu_{(i+1)k_1} \tag{5.20}$$

Note that ν_{ik_1} may be obtained via the conjugate multiplication

$$\nu_{ik_1}(n) = \arg(\hat{s}_{(i+1)k_1}(n)\hat{s}^*_{ik_1}(n)) \tag{5.21}$$

The secondary bin was defined to be one of the neighbours of the primary bin, its location can

be determined from the displacement of the partial from the centre of the primary bin

$$k_2 = \begin{cases} k_1 + 1 & \nu_{k_1} \leq 0 \\ k_1 - 1 & \text{else} \end{cases} \tag{5.22}$$

Note that the forward phase differences between the primary and secondary bins will be

$$|\nu_{ik_1} - \nu_{ik_2}| = \pi \tag{5.23}$$

giving

$$|\arg(\hat{s}_{ik_1}(n)) - \arg(\hat{s}_{ik_2}(n))| = \pi \tag{5.24}$$

To summarise, the properties of the MFT coefficients representing a partial are

1. Two non-zero coefficients (primary and secondary bins) in each time frame in which the partial is active.

2. Locally constant forward phase difference between time frames which is linearly related to the partial's frequency.

3. Constant phase difference (π) between the primary and secondary bins.

5.2.1 Partial detection

Detection of the partials is the first stage of the transcription process. At this time there is little a priori knowledge of the signal structure, due to the independent nature of the underlying score process (eqn. 4.5). Both this and the inclusion of a certain amount of noise into the signal model (eqn. 5.1) imply the use of a maximum likelyhood (ML) estimation technique for the detection of partials. The autoregressive model for this type of signal described in the previous chapter (eqn. 4.31) suggests the use of the sample covariance as a measure of the presence of a partial [Cha75]. The mth order sample covariance is defined by

$$\gamma_{mik}(n) = E\left[\hat{x}_{(i+m)k}(n)\hat{x}^*_{ik}(n)\right] \tag{5.25}$$

which may be estimated using [And71]

$$\gamma_{mik}(n) = \sum_{p,q} \hat{x}_{(p+m)q}(n)\hat{x}^*_{pq}(n) \tag{5.26}$$

However, finding suitable limits for p and q will not be possible given that there is expected to be more than one partial present in the signal as well as other features. The localisation of the partials on the time-frequency plane suggests that some windowed form of this measure will be more suitable. The local sample covariance is thus defined to be

$$c_{mik}(n) = \sum_{p,q} f_{m(i-p)} h_{m(k-q)} \hat{x}_{(p+m)q}(n)\hat{x}^*_{pq}(n) \tag{5.27}$$

where $\mathbf{f}_m$ and $\mathbf{h}_m$ are the components of a set of time-frequency separable filters.

The current system makes use of two of these measures, namely $c_{0ik}(n)$ and $c_{1ik}(n)$. The first-order covariance $c_{1ik}(n)$ gives a measure of the degree to which the transform coefficients correspond to the signal model in the region of coefficient $\hat{x}_{ik}(n)$ on MFT level n while $c_{0ik}(n)$ measures the total signal energy in that region. In order to accomplish this, it is first necessary to deduce suitable definitions for the filters $\mathbf{f}_0$, $\mathbf{h}_0$, $\mathbf{f}_1$ and $\mathbf{h}_1$. Recall from above that a partial will only contribute energy to two coefficients in any time frame. This immediately suggests that

$$h_{0k} = \begin{cases} 1 & k = 0, 1 \\ 0 & else \end{cases} \tag{5.28}$$

Similarly, but incorporating the constant phase difference in Equation (5.24)

$$h_{1k} = \begin{cases} -1^k & k = 0, 1 \\ 0 & else \end{cases} \tag{5.29}$$

Given the approximation in Equation (5.13), this definition gives measures of a partial's amplitude which are largely independent of its frequency.

Choice of $\mathbf{f}_m$ is less obvious since the extent of the partial across time in the region of $\hat{x}_{ik}(n)$ is unknown at this stage. The signal model assumes that the partial is locally stationary in this region, which suggests that a suitable definition will be some form of lowpass filter having a smooth

magnitude response with its maximum value at f_{m0}. Various forms have been investigated during the preparation of this work: the results presented in later chapters use a filter with a Gaussian response

$$f_{mk} = \begin{cases} \exp\left(\frac{-(k+0.5)^2}{2\sigma_f^2}\right) & |k| < 4\sigma_f \\ 0 & else \end{cases} \tag{5.30}$$

with $\sigma_f = 2$. This form gives a reasonable compromise between localisation of response and independence from noise.

5.2.2 Partial Onset Detection

It has been mentioned earlier that the class of signals to be analysed will contain structures other than pseudosinusoidal partials. The measures described in the previous section attempt to reject such features which do not conform to the partial model, but there may be significant energy in these features compared to some of the weaker partials and this may give rise to false results. The perceptual importance of the note onset was discussed earlier and it has been found that specific detection of these onsets can greatly improve the reliability of the partial detection by providing clues to their positions. The problem of detecting these onsets is directly analogous to the task of edge detection in image analysis which has received much attention over many years [Mar82, Jai89]. Typical amplitude behaviour of various partial onsets can be seen in the example transforms shown in figures 5.2 to 5.5. It is clear from these examples that there can be great variation in the attack rate of the partials and that this rate must be considered relative to the 'hop-size' of the MFT level on which it is observed. For the purposes of this work, a given partial onset observed on some MFT level is roughly classified as either a *'smooth'* or *'transient'* onset. Figures 5.4 and 5.5 show two transform levels of the onset of a violin note. At both resolutions, the amplitude of two active frequency bins of each partial can be seen rising smoothly out of the noise floor. This corresponds to the transient element of the signal model, (eqn. 4.18), containing negligible energy, $o(t) = 0$. Comparing these plots with figures 5.2 and 5.3, which show two similar transform levels for a piano

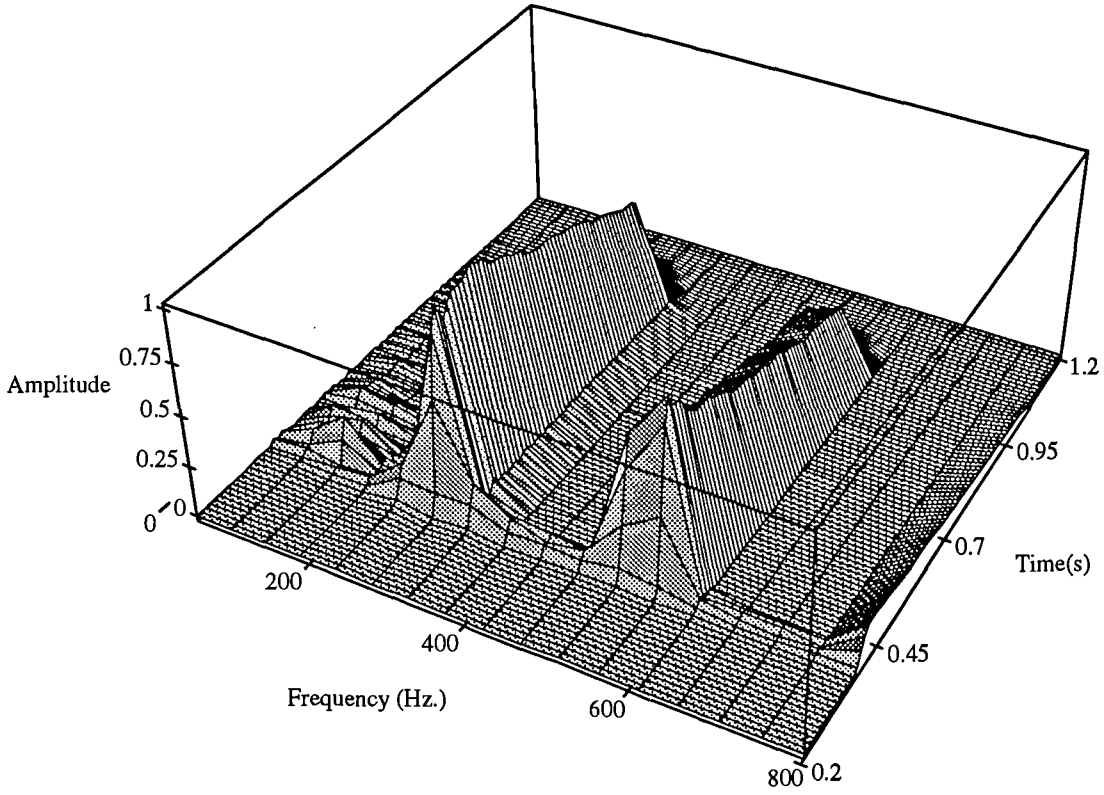

Figure 5.2: Piano note (C4) onset at MFT level 9

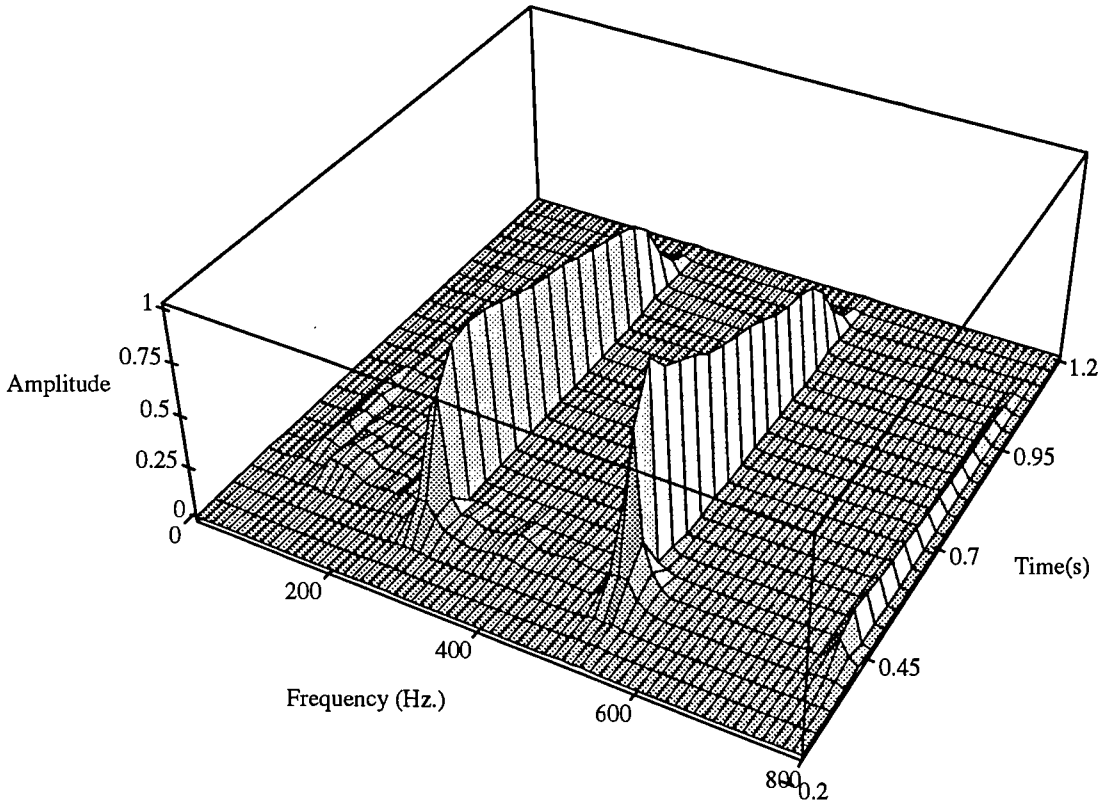

Figure 5.3: Piano note (C4) onset at MFT level 12

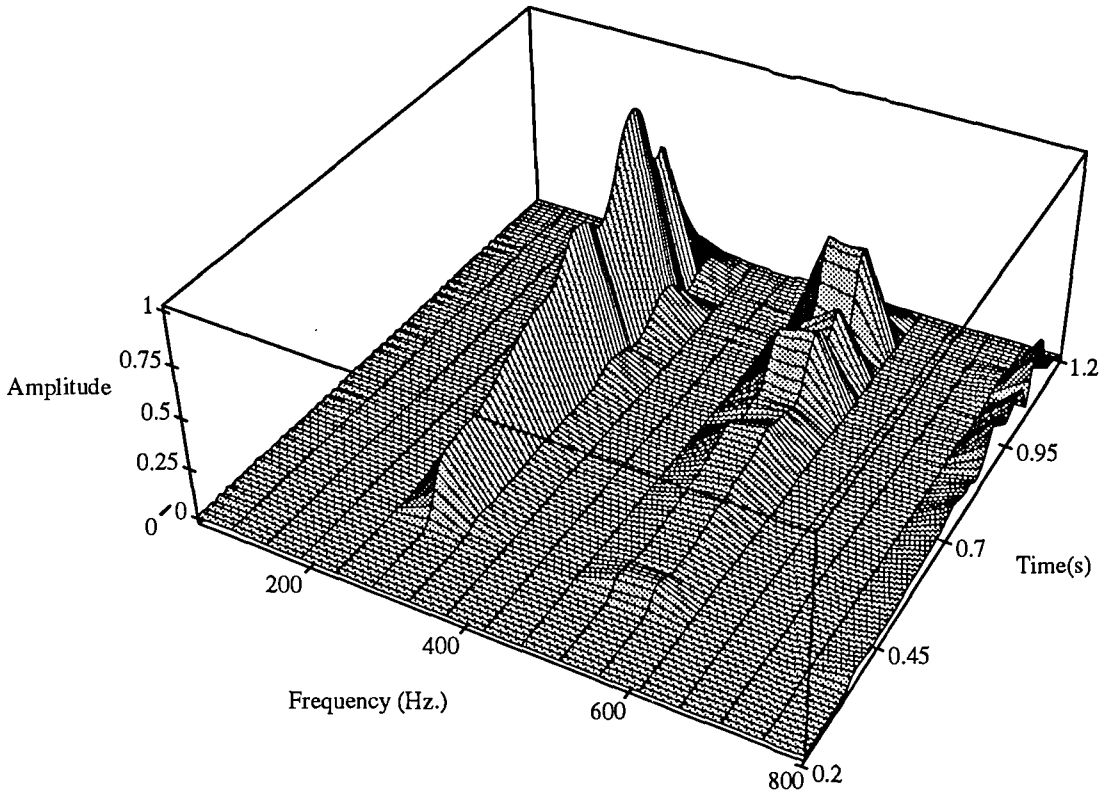

Figure 5.4: Violin note (C4) onset at MFT level 9

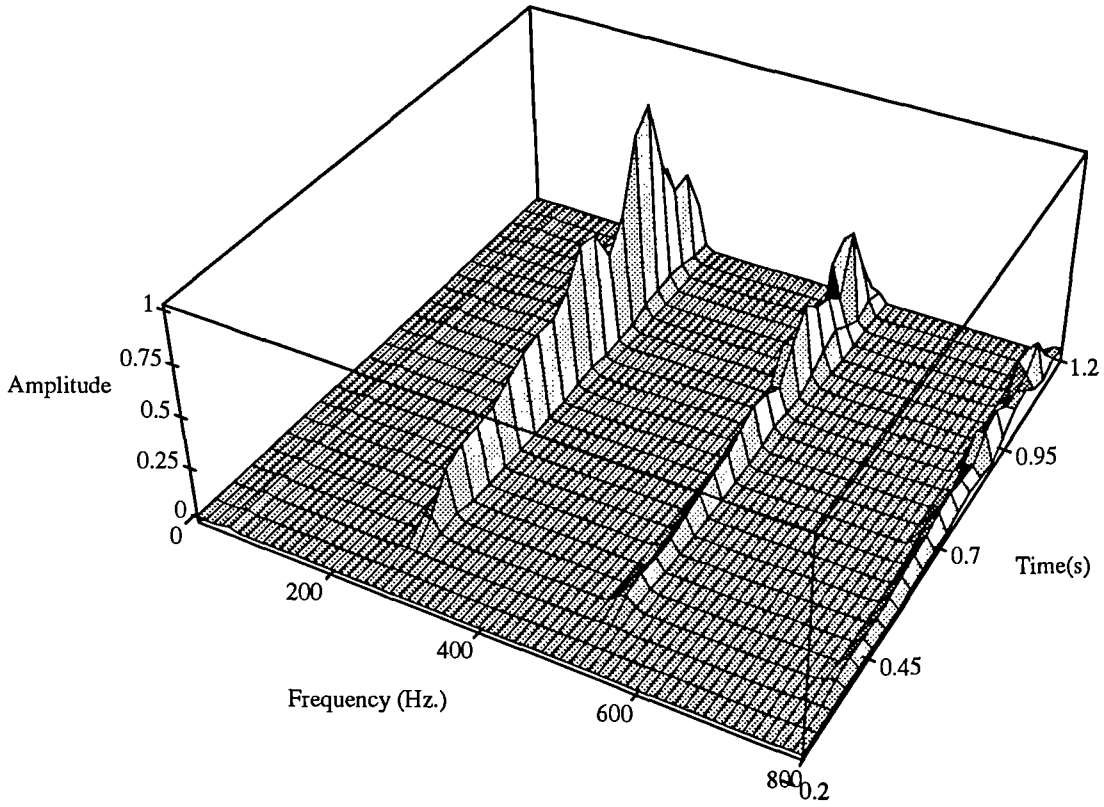

Figure 5.5: Violin note (C4) onset at MFT level 12

partial onset, reveals that the attack rate of the piano is much greater. At the higher level, which has lower temporal resolution, this results in a significant amount of energy spreading out across frequency from the centre of the partial onset. Clearly, in this case, the hypothesis that there are only ever two active coefficients for the partial fails and the detector described above will be insufficient to accurately parameterise the signal at this time. The detection of smooth onsets will be discussed first and then the technique extended to incorporate more rapid onsets.

Smooth onsets

The obvious approach to detecting onsets is to examine the gradient of the signal in each frequency bin. As an initial attempt, the time-difference between the magnitudes of each coefficient pair $\{\hat{x}_{ik}(n), \hat{x}_{(i+1)k}(n)\}$ was analysed. Unsurprisingly, this measure is highly susceptible to noise, giving a very high number of false detections. The principles of Wiener filtering (see [Pap77]) suggest that the signal model needs to be incorporated into the onset detector. The approach used currently is to analyse the time-derivative of $c_{1ik}(n)$. Since this measure is derived from more than just two coefficients it is more immune to noise in the signal and it incorporates the expected signal phase behaviour. This measure is defined in terms of a modified filter $\mathbf{f}_1'$ which has the form of the time-derivative of $\mathbf{f}_1$.

$$f'_{1k} = \begin{cases} k \exp\left(\frac{-(k+0.5)^2}{2\sigma_{f'}^2}\right) & 0 \leq k < 4\sigma_{f'} \\ -k \exp\left(\frac{-(k+0.5)^2}{2\sigma_{f'}^2}\right) & 4\sigma_{f'} < k < 0 \\ 0 & \text{else} \end{cases} \tag{5.31}$$

with $\sigma_{f'} = 2$. Since $c'_{1ik}(n)$ is required to be a measure of gradient, it is, unlike $c_{1ik}(n)$, required to be a real value. In order to preserve the sign information while removing the unknown phase, the

two lobes are calculated separately and then added after their modulus is taken

$$c'_{1ik}(n) = \left| \sum_{p=0}^{4\sigma_{f'}} \sum_{q=0}^{1} f'_{1(i-p)} h_{1(k-q)} \hat{x}_{(p+1)q}(n) \hat{x}^*_{pq}(n) \right| + \left| \sum_{p=-4\sigma_{f'}}^{-1} \sum_{q=0}^{1} f'_{1(i-p)} h_{1(k-q)} \hat{x}_{(p+1)q}(n) \hat{x}^*_{pq}(n) \right| \tag{5.32}$$

The last stage of the onset detection is to generate a set of onset events from the processed data. This operation normally includes thresholding and peak detection. It has already been noted that typical note structure includes partials with widely varying amplitudes. Clearly $c'_{1ik}(n)$ is not independent of the partial amplitude and so it will not be possible to find a threshold suitable to allow detection of partials with low amplitudes whilst avoiding false alarms caused by amplitude fluctuations in large active partials. Some amount of normalisation is required in order to obtain a measure which is more or less independent of the signal magnitude. An obvious choice for the measure to normalise $c'_{1ik}(n)$ by is $c_{0ik}(n)$ since this represents the energy in a similar region independent of the signal model. The resulting measure is thus a form of 'serial correlation' [Cha75] and is defined by

$$\rho_{ik}(n) = \frac{c'_{1ik}(n)}{\max\left[c_{0ik}(n), \varrho\right]} \tag{5.33}$$

where ϱ is a small constant selected to prevent false alarms caused by noise in regions where there is little signal energy.

Finally then, a set of onset events is obtained for each MFT level, n by thresholding $\rho_{ik}(n)$ and then finding the position of the peak value within each thresholded region. The ith onset event on level n is a vector $\theta_i(n)$ which includes the following parameters

$$\theta_{i_1}(n) = t^{\theta}_i(n) \quad \text{onset time} \tag{5.34}$$

$$\theta_{i_2}(n) = f^{\theta}_i(n) \quad \text{onset frequency} \tag{5.35}$$

$$\theta_{i_3}(n) = \delta t^{\theta}_i(n) \quad \text{onset time uncertainty} \tag{5.36}$$

$$\theta_{i_4}(n) = \delta f^{\theta}_i(n) \quad \text{onset frequency uncertainty} \tag{5.37}$$

$$\theta_{i_5}(n) = a_i^\theta(n) \qquad \text{onset certainty} \tag{5.38}$$

The position of each local maxima gives values for $t_i^\theta(n)$ and $f_i^\theta(n)$, while $a_i^\theta(n)$ is simply the value of $\rho_{ik}(n)$ at that position. The parameters $\delta t_i^\theta(n)$ and $\delta f_i^\theta(n)$ are the uncertainties (variances) of the corresponding time and frequency estimates. In the current implementation these are independent of i, being based solely on the size of the unit cells on each level.

$$\delta t_i^\theta(n) = 3\Gamma(n) \tag{5.39}$$

$$\delta f_i^\theta(n) = 2\Omega(n) \tag{5.40}$$

These values are included in each parameter vector to ease the implementation of extensions such as the detection of transient onsets described below.

The following chapter describes the process by which onset events $\theta_i(n)$ from a range of levels are combined to give some of the parameters for the partial parameter vectors $\mathbf{p}_i$.

Transient Onsets

The spreading of a partial across frequency in the region of a sudden onset, as was modelled in Chapter 4, has been incorporated in to the detection strategy. Experience has shown that this behaviour alone is not a reliable means of detecting partial onsets but is a useful means of decreasing the temporal uncertainty of detections. The reason for this was described above, often there is little or no spreading of the onset due to a slow attack rate relative to the analysis scale.

The detector used is based around a small Gaussian filter ($\sigma = 2$) straddling the partial. The filter coefficients at the primary ($k = k_1$) and secondary ($k = k_2$) bins are set to zero as the phase of the MFT coefficients in these bins does not fit the transient model (they take a far larger contribution from the steady state portion of the partial than the onset). As described in Chapter 2 (eqn. 2.64), the frequency phase difference

$$\mu_{ik}(n) = \arg(\hat{x}_{ik}(n)\hat{x}^*_{i(k-1)}) \tag{5.41}$$

is employed in the detector. A cross-frequency phase-differenced MFT coefficient is defined as

$$\tilde{x}_{ik}(n) = |\hat{x}_{ik}(n)|\, e^{-j\mu_{ik}(n)} \tag{5.42}$$

The detector uses the ratio between magnitude of the sum of these modified coefficients and the corresponding sum of the coefficient magnitudes to obtain a measure of cross-frequency phase linearity in the region of the onset.

$$\frac{\sum_{k:k\neq k_1,k_2} e^{\frac{-(k-k_1)^2}{2\sigma^2}} \tilde{x}_{ik}(n)}{\sum_{k:k\neq k_1,k_2} e^{\frac{-(k-k_1)^2}{2\sigma^2}} |\hat{x}_{ik}(n)|} \tag{5.43}$$

The closer this measure comes to unity, given sufficient magnitude in the denominator, then the more certain it is that there is an onset in this time frame i . The value of the phase difference has been found to be of little use in this application (see [Cal89]) as it is related to the centroid (eqn. 4.30) of the onset within the windowed signal, rather than its true start time; the two cannot be related without knowledge of the form of the onset.

The scheme is complicated by the fact that there may be significant energy contributions from nearby partials in the coefficients above and below the onset in frequency. Methods of avoiding this interference and the situations in which this detector is employed are discussed in the next chapter.

Chapter 6

Transcription

6.1 Introduction

The previous chapter described feature detection algorithms for some of the common structures in musical signals but did not address the question of how to integrate information from a range of MFT levels. This chapter describes how the aspects of scale and the criteria defined for determining appropriate analysis scales, discussed in general terms in Chapter 2, have been applied to the transcription of polyphonic music signals.

The detector algorithms described in the previous Chapter have been incorporated into a multiresolution detection framework such that they can operate successfully on signals containing many features. There are two basic strategies which have been used to implement this: either one level of the MFT can be selected as the most appropriate scale to apply a detector or a detector can be applied to all MFT levels and the information so produced combined across some range of levels. The actual process of combination will vary according to the nature of the data being combined and is described fully for the processing of partial onset events. Generally, however, the combination strategies are fairly independent of the features being combined and it is proposed that they could be applied to problems other than those demonstrated here. It is assumed that a set of candidate features has been

generated for each MFT level by some detection process. The simpler of the two methods is to start at some fairly high level and successively link each feature detected on that level with a corresponding feature on the level below until either the signal context suggests that correct detection on the level below would be impossible or no event can be found which satisfies the closeness of fit criterion. At this point, the chain of linked events is either, rejected, or merged into a single feature and accepted for the final set, depending on some measure of the agreement found along its length. While this technique has been found to work well it does suffer from the clear disadvantage that no feature can be in the final set that was not present in the starting set. Consequently all features must be correctly detected on the start level and this is not always possible, especially for data containing many closely grouped features. The problems involved in setting a threshold on the output of the per-level detection process were discussed in Chapter 5 (see Equation 5.33) and the lowering of thresholds required to ensure that no desired features are excluded from the initial set gives a large number of false detections due to noise. This in turn can give rise to false detections still present in the final set.

A variation on this first scheme is to start with a fairly low level and proceed upwards through the MFT's resolutions. Which variation performs the best depends on the performance of the detector used with respect to MFT level.

The second, more complex, technique is a generalisation of the first. Instead of having just one starting level, several, if not all, of the levels are considered as potential starting points for chains of linked features. Two different thresholds are used in this scheme. A higher threshold is set for the features which start a chain than for those which become linked to it. The reasoning behind this is that each feature will be best detected on the level which most closely corresponds to its natural scale but will also be detected, to some degree, on all levels where it is sufficiently isolated from neighbouring features. Setting a high initial threshold rejects all but the most certain features while the lower 'linking' threshold allows a chain of features to be built up, where such inter-level

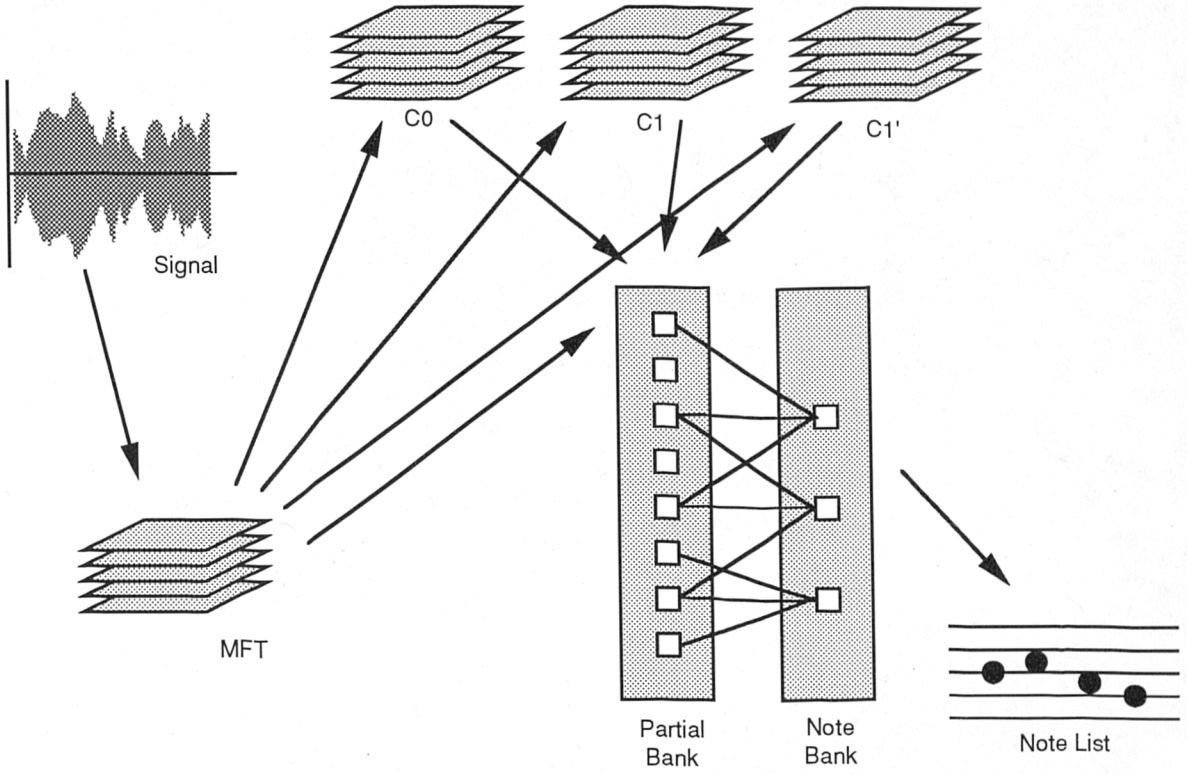

Figure 6.1: Transcription Data Structure Hierarchy

agreement exists.

Both techniques have been found to work satisfactorily, with the first being somewhat easier to implement while the second performs better in more demanding situations.

6.2 Transcription Data Structures

As was described above, determination of which levels are appropriate for the detection of some feature is dependent on the signal context and consequently much of the complexity of the current implementation is concerned with maintaining data structures to hold this contextual information. Figure 6.1 depicts the main data structures used by the transcription software. The major elements of the diagram are described in the following sections.

6.2.1 Partial Bank

At the heart of the system is the *partial-bank* structure. The most important parameter of this structure is the *partial-bank time* $t^{\mathbf{P}}$. The partial-bank holds the set of partials $\mathbf{P}$, which are active at time $t^{\mathbf{P}}$, as a frequency ordered list. Initially $t^{\mathbf{P}} = 0$ and $\mathbf{P}$ is empty. As the analysis proceeds $t^{\mathbf{P}}$ increases monotonically and partials are added and removed from $\mathbf{P}$. The use of such a scheme ensures that all calculations involved in the analysis are localised in the region of $t^{\mathbf{P}}$ and consequently it is necessary to keep those frames of the MFT in memory which represent the signal around that time. Consequently, such a scheme allows for the possibility of a real-time implementation given sufficient computational resources (see discussion of performance in Chapter 8). The current implementation keeps all MFT coefficients from levels nine to fourteen which represent the signal over the period $t^{\mathbf{P}} \pm 300$ms. in memory. In addition, the measures c_0, c_1 and c_1' are calculated for this set of coefficients and held in memory. Maintaining a list of all currently active partials enables the frequency difference between any given frequency and the nearest active partial to be easily determined. The availability of this information enables the lower bound on available MFT levels to be easily calculated.

$$\min(n) : \frac{\Delta f}{\Omega(n)} > 3 \tag{6.1}$$

i.e. the lowest level at which the two frequencies are separated by at least 3 coefficients.

6.2.2 Partials

Each partial in $\mathbf{P}$ is defined as a parameter vector $\mathbf{p}_i$, where the index, i, is unique for each partial which is a member of $\mathbf{P}$ over the duration of the analysis. The first five of the partial's parameters are scalars describing the partial's onset.

$$p_{i_1} = t_i^p \qquad \text{partial onset time} \tag{6.2}$$

$$p_{i_2} = \delta t_i^p \qquad \text{partial onset time uncertainty} \tag{6.3}$$

$$p_{i_3} = f_i^p \quad \text{partial onset frequency} \tag{6.4}$$

$$p_{i_4} = \delta f_i^p \quad \text{partial onset frequency uncertainty} \tag{6.5}$$

$$p_{i_5} = a_i^p \quad \text{partial onset magnitude} \tag{6.6}$$

The next parameter is a list of parameter vectors representing the time-frequency path of the partial

$$\mathbf{p}_{i_6} = \mathbf{q}_i \tag{6.7}$$

Each element of $\mathbf{q}_i$, $\mathbf{q}_{ik}$ has four elements characterising a point on the path of the partial

$$q_{ik_1} = t_{ik}^q \quad \text{time} \tag{6.8}$$

$$q_{ik_2} = f_{ik}^q \quad \text{frequency} \tag{6.9}$$

$$q_{ik_3} = a_{ik}^q \quad \text{amplitude} \tag{6.10}$$

$$q_{ik_4} = l_{ik}^q \quad \text{MFT level} \tag{6.11}$$

$\mathbf{q}_i$ is kept in time-order such that

$$t_{ik_n} < t_{ik_m} \quad \text{where} \quad n < m \tag{6.12}$$

In addition to the time, frequency and amplitude being recorded for each point on the partial, notice that the MFT level is also stored; partials are allowed to change their scale from frame to frame. Keeping a record of the level is not strictly necessary to describe the partial's path but has proved useful for display and debugging purposes and would possibly be required by some subsequent processing stage.

For notational convenience, the most recently added values of amplitude and frequency to the path vector are defined as:

$$\mathcal{F}_i^p \quad \text{current partial frequency} \tag{6.13}$$

$$\mathcal{A}_i^p \quad \text{current partial amplitude} \tag{6.14}$$

The final parameter associated with each partial is a vector of references to notes in the note-bank

$$\mathbf{p}_{i_7} = \mathbf{h}_i^p \tag{6.15}$$

6.2.3 Note Bank

The *Note-bank* structure is similar to the partial-bank. It is a container for the set, **N**, of current note hypotheses which are held as a list in order of increasing onset frequency. The note-bank is subordinate to the partial-bank in the sense that **N** is the set of hypotheses at the partial-bank time $t^{\mathbf{P}}$.

6.2.4 Note Hypotheses

Each note hypothesis, $\mathbf{n}_i$, is a potential note. During the course of the transcription the existence of many notes is hypothesised; more information becomes available as the partial and note banks progress through time and, in the light of this, many of the hypotheses are rejected and removed from **N**. As with partials, the index, i, is over all note hypotheses which are members of **N** for any period during the analysis. $\mathbf{n}_i$ contains the following scalar parameters:

$$n_{i_1} = t_i^n \quad \text{note onset time} \tag{6.16}$$

$$n_{i_2} = f_i^n \quad \text{note onset fundamental frequency} \tag{6.17}$$

as well as two vectors

$$\mathbf{n}_{i_3} = \mathbf{m}_i \quad \text{note fundamental path} \tag{6.18}$$

$$\mathbf{n}_{i_4} = \mathbf{h}_i^n \quad \text{note's partials} \tag{6.19}$$

The path vector $\mathbf{m}_i$ is a record of the estimated fundamental frequency of and total energy belonging to the note with time. It is similar to the path vectors of partials, with each element being a parameter

vector having three elements

$$m_{ik_1} = t_{ik}^m \quad \text{time} \tag{6.20}$$

$$m_{ik_2} = f_{ik}^m \quad \text{frequency} \tag{6.21}$$

$$m_{ik_3} = a_{ik}^m \quad \text{amplitude} \tag{6.22}$$

The last note parameter, h_i^n, is a vector of references to partials in the partial-bank. Each element, h_{ik}^n, where $1 \leq k < N_{max}^h$, is the index of a partial which is hypothesised as being the kth harmonic of note i. This linking vector is used by the note-partial linking algorithm to associate a set of partials with each note. The current implementation links up to ten partials with each note, $N_{max}^h = 10$. As was mentioned above, there is a corresponding linking vector of note indices associated with each of the partials which is used by the same algorithm. In this case each element is h_{ik}^p, again with $1 \leq k < N_{max}^h$, and indicates the note which is considering partial i as its kth harmonic at that time. The use of this double linking enables the mapping between notes and partials to be easily changed. No partial in the partial-bank or note hypothesis in the note-bank has an index of zero. This value is used to indicate an empty element in the linking vectors.

6.2.5 Accepted Note List

This list stores notes which have been removed from the note-bank. These notes have not been rejected, but have been determined to have ended and so should not be present in the note-bank. Although data on the performance of the transcription system is derived from many of the data-structures, this list can be considered to be the primary set of results.

6.3 Transcription Algorithms

6.3.1 Generating Onset Events

The method of generating candidate onset events from the correlation measures derived from each MFT coefficient was described in the preceding chapter. The first task of the transcription system is to ascertain the positions of the partial onsets; this is done by combining these candidate events across scale. Such a process results in a set of onset events which have decreased time and frequency uncertainties and far fewer false-alarms than the initial sets. Both the generation of the candidate events and their combination are performed in advance of the partial-bank by an amount which assures that all combination and revision processing for a given partial onset will have been completed by the time the partial-bank arrives at that point. However, as will be seen below, such computations make use of the positions of the partials within P, and so it is necessary to make the assumption that these positions will remain fairly constant over this period. The current implementation sets this *bank offset time* to 200ms. This is largely determined by the duration of the coefficients in the highest MFT level used. The offset could be shortened for most events by the use of a more complex algorithm which would take into account the current level of the nearest partials rather than using the worst case, the highest level.

The second scale combination algorithm, described above, is used for combining the onset events. The two thresholds used are A_l^θ and A_h^θ with $A_l^\theta < A_h^\theta$. Starting at each level, events which have a magnitude $a_i^\theta(n)$ greater than A_h^θ are used as the starting point of a linked chain of events. The 'distance' measure, d^θ, between two events $\theta_1(n)$ and $\theta_2(n-1)$ on neighbouring levels n and $n-1$ is calculated using

$$d_{1,2}^\theta = \frac{\left|(t_1^\theta(n) - t_2^\theta(n-1)\right|^2}{\delta t_1^\theta(n) + \delta t_2^\theta(n-1)} + \frac{\left|(f_1^\theta(n) - f_2^\theta(n-1)\right|^2}{\delta f_1^\theta(n) + \delta f_2^\theta(n-1)} \tag{6.23}$$

A chain which already includes $\theta_1(n)$ is extended by selecting $\theta_2(n-1)$ from the events on level $n-1$ with $a_2^\theta(n-1) \geq A_l^\theta$ so as to minimise $d_{1,2}^\theta$. However $\theta_2(n-1)$ is only added to the chain if both

the conditions

$$\left|t_1^\theta(n) - t_2^\theta(n-1)\right|^2 < \delta t_1^\theta(n) + \delta t_2^\theta(n-1) \tag{6.24}$$

$$\left|(f_1^\theta(n) - f_2^\theta(n-1)\right|^2 < \delta f_1^\theta(n) + \delta f_2^\theta(n-1) \tag{6.25}$$

are satisfied. If the chain cannot be extended to this level then the extension process terminates, since the chains must be contiguous. An event chain is merged and accepted as a partial onset if it contains more than one event. The resulting combined event, $\theta_{k'}$, has the same parameter list as the level events. These are calculated by successively combining events $\theta_k(n)$ and $\theta_j(n-1)$ along the chain using the following formulae which are based on the assumption that the estimation errors made at each level are independent.

$$t_{k'}^\theta = \frac{t_k^\theta \delta t_j^\theta + t_j^\theta \delta t_k^\theta}{\delta t_j^\theta + \delta t_k^\theta} \tag{6.26}$$

$$f_{k'}^\theta = \frac{f_k^\theta \delta f_j^\theta + f_j^\theta \delta f_k^\theta}{\delta f_j^\theta + \delta f_k^\theta} \tag{6.27}$$

$$\delta t_{k'}^\theta = \frac{2\delta t_j^\theta \delta t_k^\theta}{\delta t_j^\theta + \delta t_k^\theta} \tag{6.28}$$

$$\delta f_{k'}^\theta = \frac{2\delta f_j^\theta \delta f_k^\theta}{\delta f_j^\theta + \delta f_k^\theta} \tag{6.29}$$

$$a_{k'}^\theta = \max\left(a_k^\theta, a_j^\theta\right) \tag{6.30}$$

At this point in the process, the transient onset detector, described at the end of Chapter 5 (eqn. 5.43) is employed to decrease the temporal uncertainty of the onset estimates. The detector is applied at each point where an onset has been found, as directed by the preceding process, at the highest MFT level employed in this detection. If the transient detector indicates that the cross-frequency differenced phase is approximately linear in that region then the temporal uncertainty of

the onset is reduced to the duration of that time frame if this is less than the uncertainty already obtained by the cross-level combination process.

Each of the accepted events are used to initiate a new partial $\mathbf{p}_i$. The first five partial parameters are simply copies of the corresponding onset event parameters.

$$t_i^p = t_{k'}^\theta \tag{6.31}$$

$$\delta t_i^p = \delta t_{k'}^\theta \tag{6.32}$$

$$f_i^p = f_{k'}^\theta \tag{6.33}$$

$$\delta f_i^p = \delta f_{k'}^\theta \tag{6.34}$$

$$a_i^p = \theta_{k'} \tag{6.35}$$

The change in index is required since many candidate onset events have been rejected by this stage and it is simpler to keep the indices of the partials contiguous. Thus $\mathbf{p}_i$ is the ith partial to be added to the partial-bank set $\mathbf{P}$. It should be noted that, due to the bank offset time, the onset time of this partial will be in advance of the partial-bank time ($t_i^p > t^{\mathbf{P}}$) and so the new partial must remain dormant within the bank until $t_i^p = t^{\mathbf{P}}$. This dormant phase is in fact the first of three that members of $\mathbf{P}$ may pass through; the phases are

Dormant awaiting bank to arrive at its onset time

Active keeping up with bank, seeking path through the MFT data

Dead no longer active ($t_i^p < t^{\mathbf{P}}$) awaiting removal from the bank.

The Active and Dead phases are discussed below.

6.3.2 Forming Note Hypotheses

At the same time a new partial $\mathbf{p}_i$ is added to the partial-bank a corresponding set of note hypotheses is added to the note-bank. These entities represent the hypotheses that the new partial is the jth

harmonic of some note; with $1 \leq j \leq N^H_{max}$. For each j the note hypothesis' parameters are initialised such that

$$t^n_{k+j} = t^p_i \tag{6.36}$$

$$f^n_{k+j} = f^p_i / j \tag{6.37}$$

$$h^n_{(k+j)j} = i \tag{6.38}$$

given that the last hypothesis to be added to **N** was the kth. Additionally the links in the partial are set up to link to the created note hypotheses

$$h^p_{ij} = k + j \tag{6.39}$$

The choice of a suitable value for N^H_{max} has been the subject of some experimentation. It has been found not unreasonable to constrain the system such that $N^H_{max} = 1$, i.e. generating just one note hypothesis for each new partial such that the partial is its first harmonic. This produces good results in most circumstances, only failing when the first harmonic of the note cannot be detected, a so called *missing fundamental* [Sma70]. No such cases have been found for the set of natural instruments analysed for this work, though such signals can be easily generated.

Each note hypothesis may pass through several stages during its lifetime. Possible stages are *linking, mature, defunct* and *dead.* Newly generated hypotheses are initially in the linking phase.

6.3.3 Note-Partial linking

The note-partial linking algorithm is executed once during every main loop of the transcription algorithm i.e. once for every time frame of the lowest MFT LEVEL. It acts to link note hypotheses in the note-bank with a set of partials from the partial-bank which is best suited to be the note's harmonics. The algorithm runs as follows: for each note n_i in the bank which is in its linking phase an initial estimate of its fundamental frequency f^n_i is made from the lowest currently linked

harmonic, $h^n_{il_{min}}$.

$$l_{min} = \min(l) : h^n_{il} \neq 0 \tag{6.40}$$

$$\hat{f}^n_i = \frac{1}{l_{min}} f^p_{h^n_{il_{min}}} \tag{6.41}$$

Then, for each harmonic, l, in turn, its expected frequency is estimated by adding the note's fundamental frequency onto the frequency of the previous harmonic or it's estimate

$$\hat{f}^p_{h^n_{il}} = \begin{cases} f^p_{h^n_{i(l-1)}} + f^n_i & h^n_{i(l-1)} \neq 0 \\ \hat{f}^p_{h^n_{i(l-1)}} + f^n_i & \text{else} \end{cases} \tag{6.42}$$

Using this estimate, the partial-bank is searched for the partial nearest to this frequency with a similar start-time. If a suitable candidate is found and it is a better match, in terms of onset time and frequency, than any previously linked partial then the linking vectors are adjusted to link this partial and note. The note's fundamental frequency is then re-estimated using the partials linked so far

$$f^n_i = \frac{\sum_{j:h^n_j \neq 0}^{l} f^p_{h^n_{ij}}}{\sum_{j:h^n_j \neq 0}^{l} j} \tag{6.43}$$

Recalculation of the fundamental for each harmonic avoids cumulative errors building up in the estimated frequencies of the upper harmonics. The process continues until either all harmonics have been processed or the number of unlinked harmonics exceeds three. The note start-time is then revised according to the start-times of the linked partials

$$t^n_i = \frac{\sum_{j:h^n_j \neq 0}^{l} t^p_{h^n_{ij}}(l - j + 1)}{\sum_{j:h^n_j \neq 0}^{l} j} \tag{6.44}$$

Using this algorithm, each note hypothesis seeks out the set of partials which best fits its harmonic frequencies. Running the algorithm in every time-frame allows the chosen set to change freely during the early stages of a note so avoiding problems caused by the lack of onset synchronisation observed

between the harmonics of many notes. The algorithm is not run on notes which have passed out of the linking phase.

6.3.4 Partial-bank iteration

The partial-bank iteration algorithm is executed for every time-frame. The first stage is to scan the list of partials to see which of the dormant partials are to become active on this frame. Those are then allocated a scale using Equation (6.1) and marked as being in their active phase. This part of the transcription system can suffer from problems. The idea of having dormant partials in the partial-bank for some time before becoming active is to allow the already active neighbouring partials to adjust their scale in order to avoid a collision. However, when a dormant partial is very close in frequency to one of the other members of the bank there may be no available scale with sufficient frequency resolution to separate them. In such cases, the two partials are probably coincident and would certainly be perceived as one tone by a human listener even if they were attributable to separate source instruments. Clearly, trying to separate them by means of an MFT level with excessively high frequency resolution is not a desirable strategy; the current approach in such situations is to arbitrate between the two partials, removing one of them from the bank. A very simple method of choosing which one to remove is used; if the already active partial is older than 50ms, then it is removed; otherwise the new partial is removed. A better scheme is suggested in the final chapter.

On the next pass along the bank each partial is given the opportunity to advance; whether it does or not depends on its current level, since the partial-bank iteration algorithm is run for every frame in the lowest level. When a partial advances it moves forward in time by the duration of one frame on its current level and records this new position in its path vector. Thus for partial i with k elements already in its path vector $\mathbf{q}_i$, the next time is calculated using

$$t^q_{i(k+1)} = t^q_{ik} + \Gamma(l^q_{ik}) \tag{6.45}$$

The frequency of the partial is allowed to vary in order to track the frequency variations of the note. There are two parts to this process: a coarse variation is accomplished by following the peak in the amplitude of the measure $c_1(n)$ with the restriction that the partial may not move more than one frequency bin per time frame. The phase of the selected coefficient is then used to determine the frequency of the partial using the method shown in Equation (5.19) but using $c_1(n)$ rather than the forward phase difference of the level coefficients directly. This gives an estimate with increased noise immunity. The path frequency is recorded in $f^q_{i(k+1)}$. Lastly, the path amplitude of the partial $a^q_{i(k+1)}$ is calculated from the primary and secondary level coefficients (Equation 5.13).

As each partial advances it is also given the chance to change level. The current level of each partial is the lowest available level which satisfies Equation (6.1) using the frequency difference between itself and its closest neighbour. Moving up MFT levels is accomplished by choosing a coefficient on the destination level which corresponds to the current time-frequency position of the partial. The current partial frequency is, however, not known with sufficient certainty to determine the corresponding frequency bin on the higher level. This problem is overcome by choosing the coefficient in that area with the largest amplitude. Moving down levels is relatively simple, the chosen coefficient is that just proceeding the temporal mid-point of the current coefficient.

The third and final pass over the partial list determines whether any of the partials should be terminated. This decision is based on two criteria: either the absolute magnitude of the partial has remained below -80dB [1]. for more than 50ms. or the value of $c_0(n)$ at the current coefficient is less than 1% of its maximum value during the lifetime of the partial. These criteria were derived by experimentation. Partials which are terminated pass into the dead phase for later removal from the list.

[1] the original signal is normalised during the generation of the MFT so that the largest signal peak has a mgnitude of 0dB.

6.3.5 Note-bank iteration

The note-bank, as well as the partial-bank, has an iteration algorithm which is called for every time-frame on the lowest MFT level. The algorithm's purpose is to update the parameters and phases of all of the note hypotheses in the note-bank. The note hypothesis phases are a little more complex than the partial phases and are defined as follows:

Linking. The note-partial linking algorithm is run for each note hypothesis recently introduced to the bank on every time-frame in order to establish the best fitting set of partials.

Mature. In order to become mature, a note hypothesis must satisfy certain conditions, described below, on the quality of fit of its partials. Mature notes cannot change the linking of their partials. Both the linking and mature phases are active phases but this phase can be considered to be the normal phase for a successful note hypothesis.

Defunct. A defunct note hypothesis has been rejected for some reason and will never be transferred to the accepted note list. After entering this phase the note hypothesis becomes inactive and is removed from the bank on the next iteration.

Dead. When a mature note hypothesis has been determined to have finished then it passes to the dead phase. It becomes inactive and is transferred from the bank to the list of accepted notes on the next iteration.

The processes by which note hypotheses are formed, added to the note-bank and behave during the linking phase have been described above. The number of note hypotheses generated and added to the bank is much larger than that expected to be accepted as notes since at least one hypothesis is generated for each partial detected. The number of hypotheses on the note-bank is reduced by considering the quality of fit of the set of partials which become associated with each hypothesis during its linking phase. In the current system, this calculation is performed on hypotheses after

they have been in the bank for longer than 80ms. Such a fixed evaluation time limits the minimum length of note which can be detected by the system, though this value has been found suitable for the test data analysed in the following chapter; a better algorithm is suggested in the final chapter.

The quality of fit is a correlation measure and is defined on each note hypothesis $\mathbf{n}_i$ as the sum of the products of the normalised partial onset certainties of adjacent pairs of its partials

$$r_i = \sum_{j=2}^{N^H_{max}} a^p_{h^n_{i(j-1)}} a^p_{h^n_{ij}} \tag{6.46}$$

$a^p_{h^n_{ij}}$ being the onset certainty a^p for the jth partial of the ith note.

This measure is derived from the proposition that the normalised onset magnitudes of the set of partials belonging to a note may be modelled using a Markov model

$$a^p_{h^n_{ij}} = \gamma a^p_{h^n_{i(j-1)}} + \epsilon_{ij} \quad j > 1 \tag{6.47}$$

where $0 < \gamma < 1$. ϵ_{ij} are random variables with mean $(1-\gamma)a^p_{h^n_{i1}}$.

The note hypotheses are reduced by simply discarding (making defunct) any which have a quality measure below a certain threshold, $r_i < A^r$. Such a simple technique does not yield perfect results, the aim is merely to remove the most unlikely hypotheses, any further reduction would almost certainly require algorithms which incorporate some amount of stylistic and contextual knowledge. As was discussed earlier, such algorithms are outside the scope of this work. Even so, this stage of the processing reduces the number of hypotheses by 60-80% and rarely removes any which correspond to actual notes in the original data.

The note bank iteration algorithm, in addition to updating the phases of the note hypotheses using the methods described above, calculates the parameters stored on the path vector for each hypothesis in either of the linking or mature phases. These calculations are performed for every time frame resulting in a detailed description of the fundamental path for each note. For the ith note, already having $j-1$ elements on its path vector, the jth member has its three elements calculated

using the following formulae

$$(a_{ik}^m)^2 = \sum_{l=1:h_{il}^n \neq 0}^{N_{max}^h} \left(\mathcal{A}_{h_{il}^n}^p\right)^2 \tag{6.48}$$

i.e. the amplitude of the note is simply the sum of the amplitudes of its partials. The fundamental frequency is calculated as the mean of the partial frequencies divided by their partial number and weighted by their amplitude

$$f_{ik}^m = \frac{\sum_{l=1:h_{il}^n \neq 0}^{N_{max}^h} \mathcal{A}_{h_{il}^n}^p \mathcal{F}_{h_{il}^n}^p / l}{\sum_{l=1:h_{il}^n \neq 0}^{N_{max}^h} \mathcal{A}_{h_{il}^n}^p} \tag{6.49}$$

Lastly, the note's path vector time is set to the partial-bank time

$$t_{ik}^m = t^{\mathbf{P}} \tag{6.50}$$

The final action of the note-bank iteration algorithm is to identify notes which have terminated and mark them as dead. In the current implementation, notes are deemed to have terminated when all their harmonics have terminated.

At regular intervals, the note-bank is searched for dead notes. These are then removed from the bank and transferred to the final note list.

Chapter 7

Results

7.1 Introduction

This chapter presents some results obtained using the algorithms presented in the preceding chapters. First there is an example of the use of phase differencing for pitch detection. Next the representations at all stages of a full transcription of the two piano note signal, seen in Chapter 1, are shown. The analysis of a short section of a Bach woodwind trio then demonstrates the ability of the system to cope with a signal containing many features. Finally, a short piece of a Schubert trio is analysed and shows some of the shortcomings of the present system.

The analyses were performed on a Sun Microsystems 4/480 SPARC based computer with 32Mb. of memory. The time taken was approximately 1.5 – 2 minutes for each second of audio; 90in the generation of the MFTs themselves. The use of more specialised signal processing hardware and computing each MFT level in parallel would significantly reduce these times.

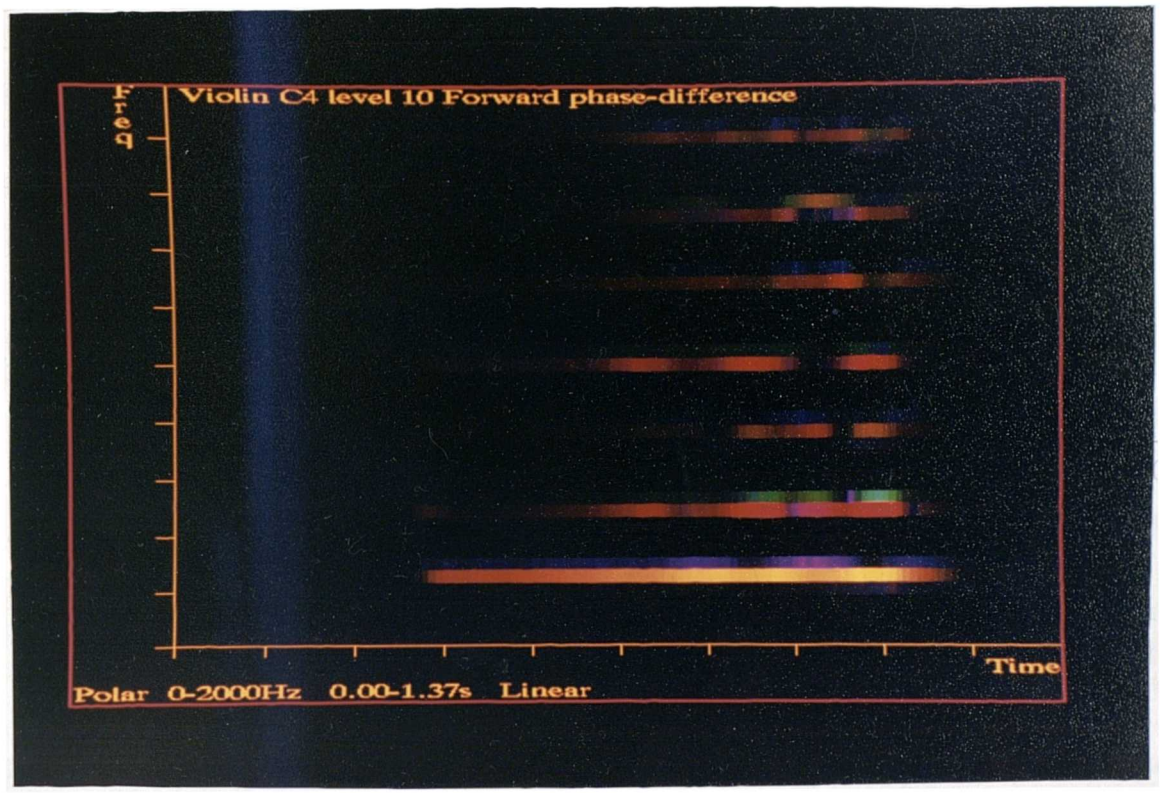

Figure 7.1: Violin note showing forward phase-difference

7.2 Phase Differencing

The use of the differenced phase to detect the pitch and onset time of partials was discussed in chapters 3 and 4. Figure 7.1 shows level 9 of an MFT of a single violin note C4. This plot is different from the preceding plots in that the colour of each coefficient indicates its complex value. As before, the brightness at each point represents the magnitude of the coefficient, but the hue indicates the phase. The mapping between hue and phase has been chosen so that a smooth variation of phase is seen as a smooth variation of hue. The absolute value of this phase is not important for the results shown here. The phase of the violin MFT level has been differenced across time. Notice that, for the first half of the note, the colour of each of the two active coefficients is constant. After that point a gentle vibrato is introduced; this can be seen as variation of colour along the partial, accompanied by a shifting in the distribution of partial energy between adjacent frequency bins. The effect is more noticeable in the upper harmonics, where the frequency deviation is larger.

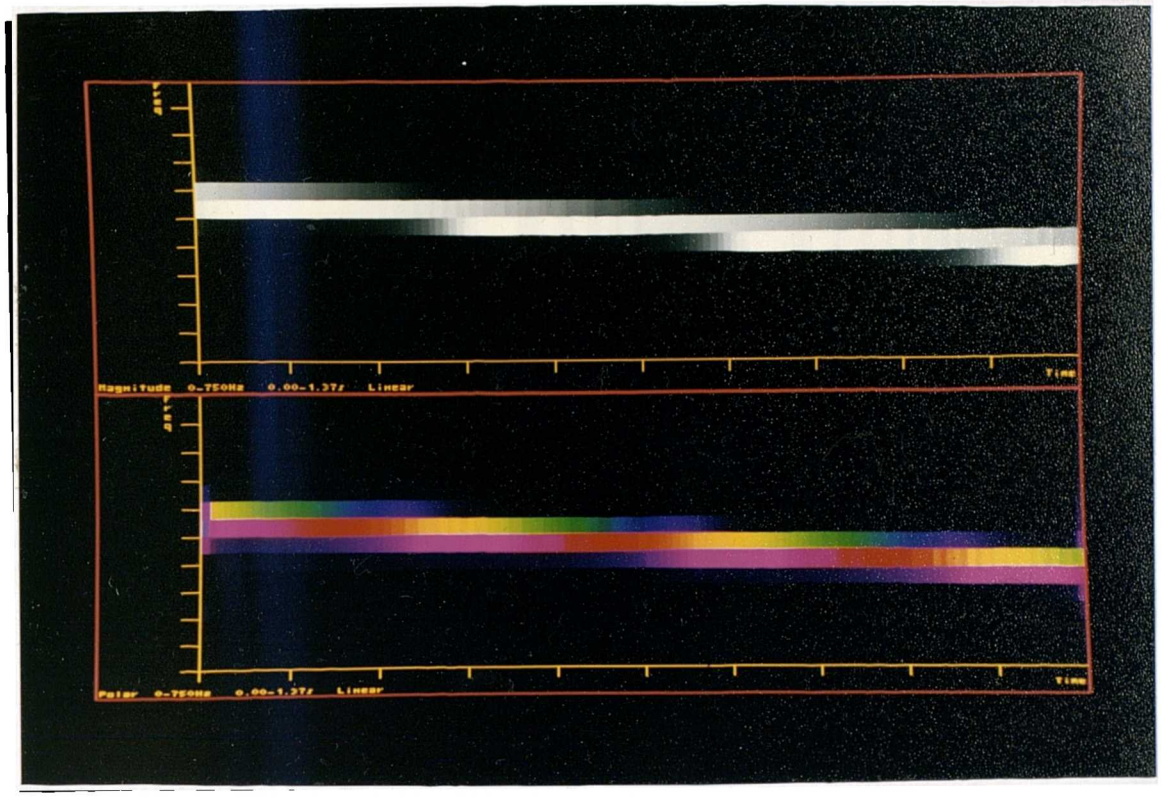

Figure 7.2: Polar and magnitude only plots of tone with changing frequency

The effect of frequency variation on the forward phase difference can be more clearly seen in Figure 7.2 which shows two plots of level 10 of the MFT of a single tone of changing frequency. The tone starts at 400 Hz and decreases in frequency by 200 Hz over 1.37 seconds. The upper plot shows just the magnitude of the level coefficients. Note the way energy moves smoothly between frequency bins using this (relaxed) version of the MFT. There is some rapid variation in magnitude with time, which can be attributed to beating caused by interference due to the temporal sidelobes of the FPSS basis function. The lower plot is a polar display of the forward differenced phase for the same level. The colour changes smoothly with time, aside from some beating, indicating the proportional relationship between partial frequency and the time-differenced phase. There is a constant phase difference between adjacent bins of π, while the phase in the primary bin, that with the largest magnitude in any time frame, is always within the range $\pm\frac{\pi}{2}$. These relationships enable the frequency of the tone to be determined unambiguously from the forward-differenced phase of

the primary bin, as has been described above (eqn. 5.19).

7.3 Two Piano Notes

This section examines, step by step, the analysis of the two piano note signal which was first discussed in Chapter 1. The signal consists of F♯ above middle C followed by middle C. The sound was assembled in the digital domain from the McGill University Master Samples compact disc [OW87] Volume 3, Track 2, Bands 46 and 40, by simple addition. The relative magnitudes of the two notes were left as they were on the disc while the onsets of the notes were set at 1.18 s and 1.38 s, as could best be judged using a graphical signal editor. The onset in this case is taken to be the point at which the sample magnitude is observed to start rising above the noise floor, which was measured at -70dB for this signal. In what way this definition of the onset time is at variance with the perceived onset time has not been investigated in this work, due to its non-clinical nature. It is assumed that the perceived onsets would be at somewhat later times. The sound has been truncated at about 2 seconds using a rectangular window, all sample magnitudes after this point being set to zero.

The signal was first transformed using a MFT incorporating oversampling by a factor of two. Five levels, (9-13), were calculated and are discussed in the next section.

7.3.1 The MFT Levels

Figures 7.3 to 7.7 display all the computed MFT levels for the two piano note signal. The amplitude scales of these plots is a logarithmic (dB) scale with the black level of the plots being set at -50dB below a white level of 0dB. Such a scale is effective in revealing, somewhat disproportionately, the low magnitude details present. For instance, it is evident from levels 10 and 11 that there is a fairly large amount of low frequency noise at the time of the note onsets. The next most obvious feature revealed is the effect of the analysis window's sidelobes in the time domain. This is clearly shown

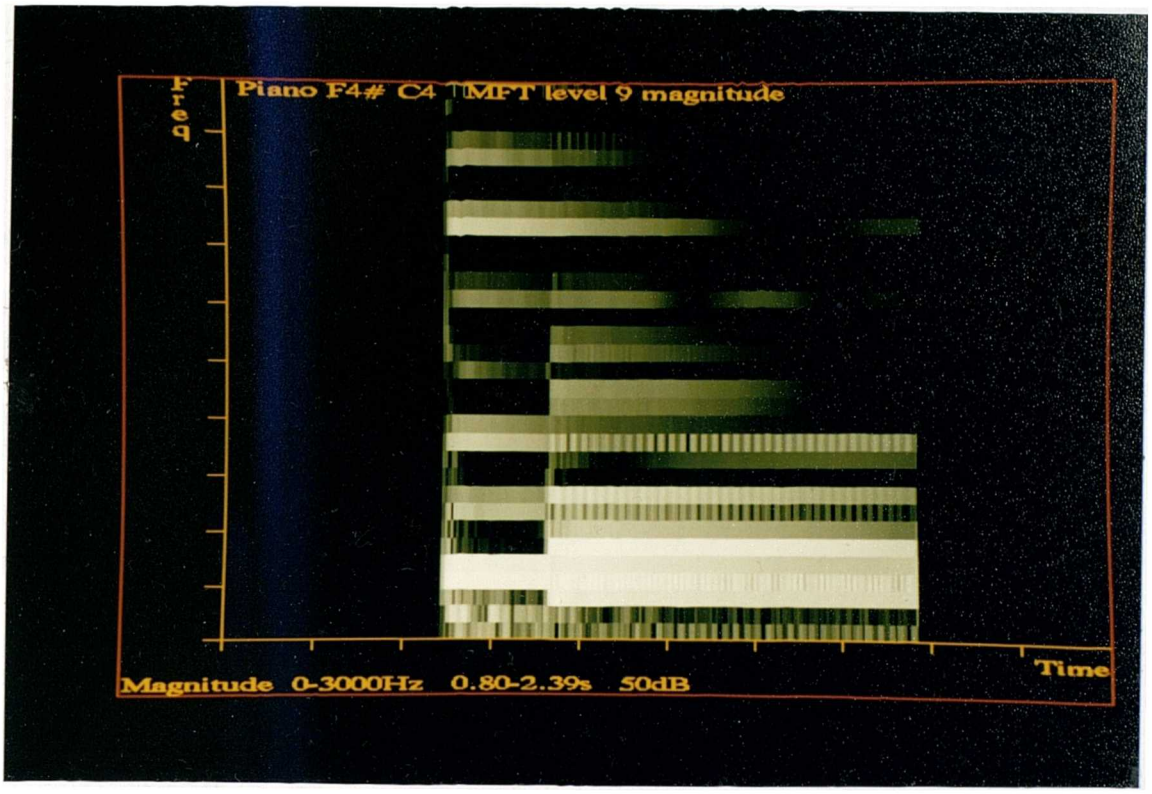

Figure 7.3: Two Piano Notes: MFT level 9, magnitude

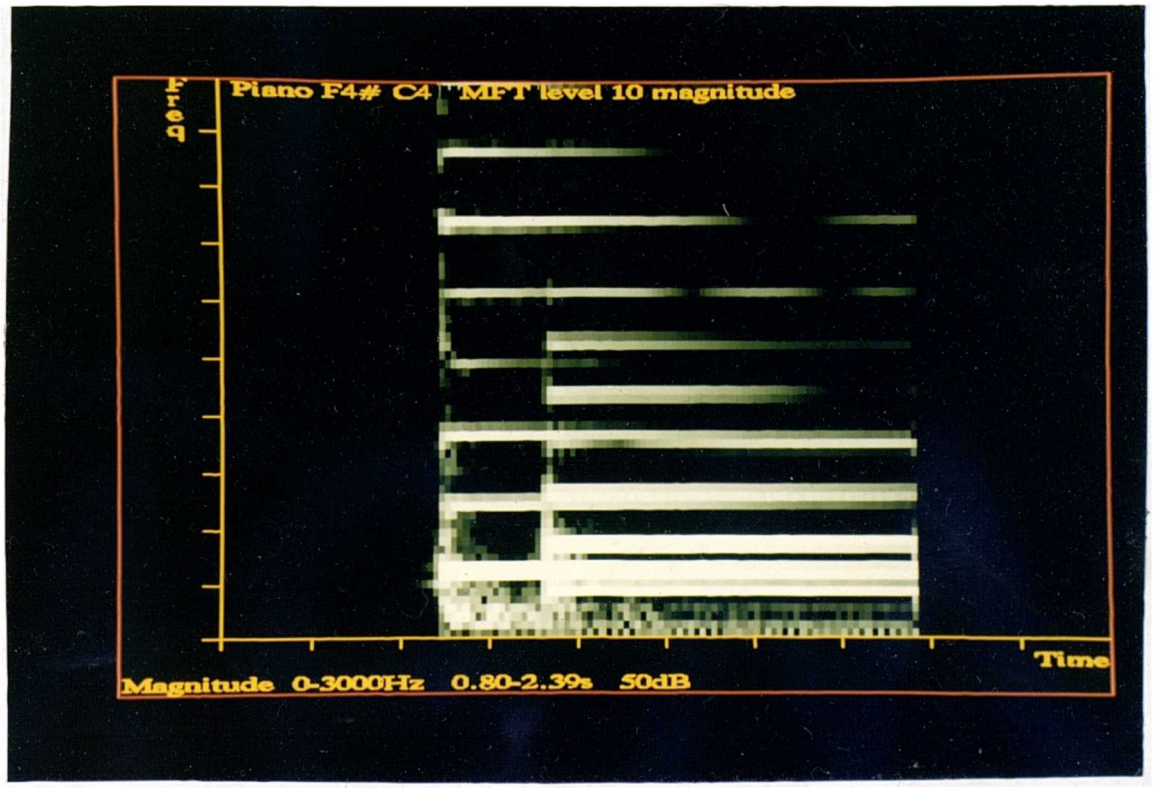

Figure 7.4: Two Piano Notes: MFT level 10, magnitude

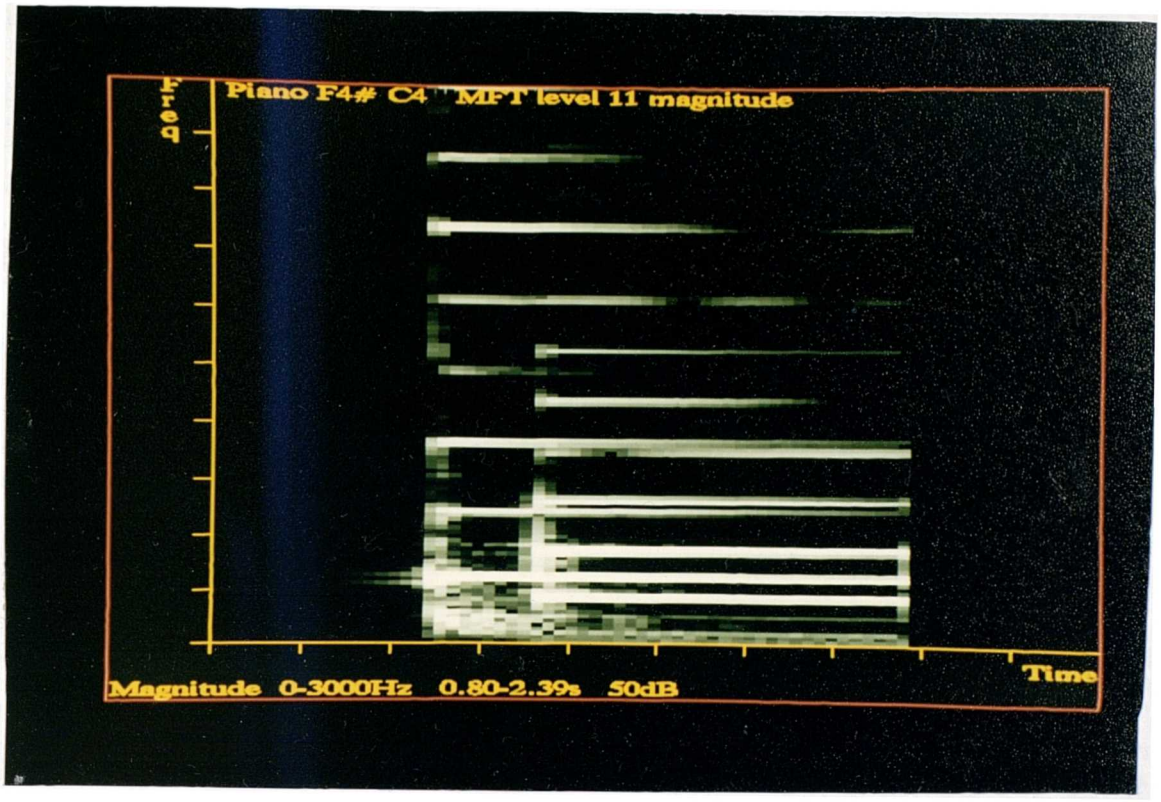

Figure 7.5: Two Piano Notes: MFT level 11, magnitude

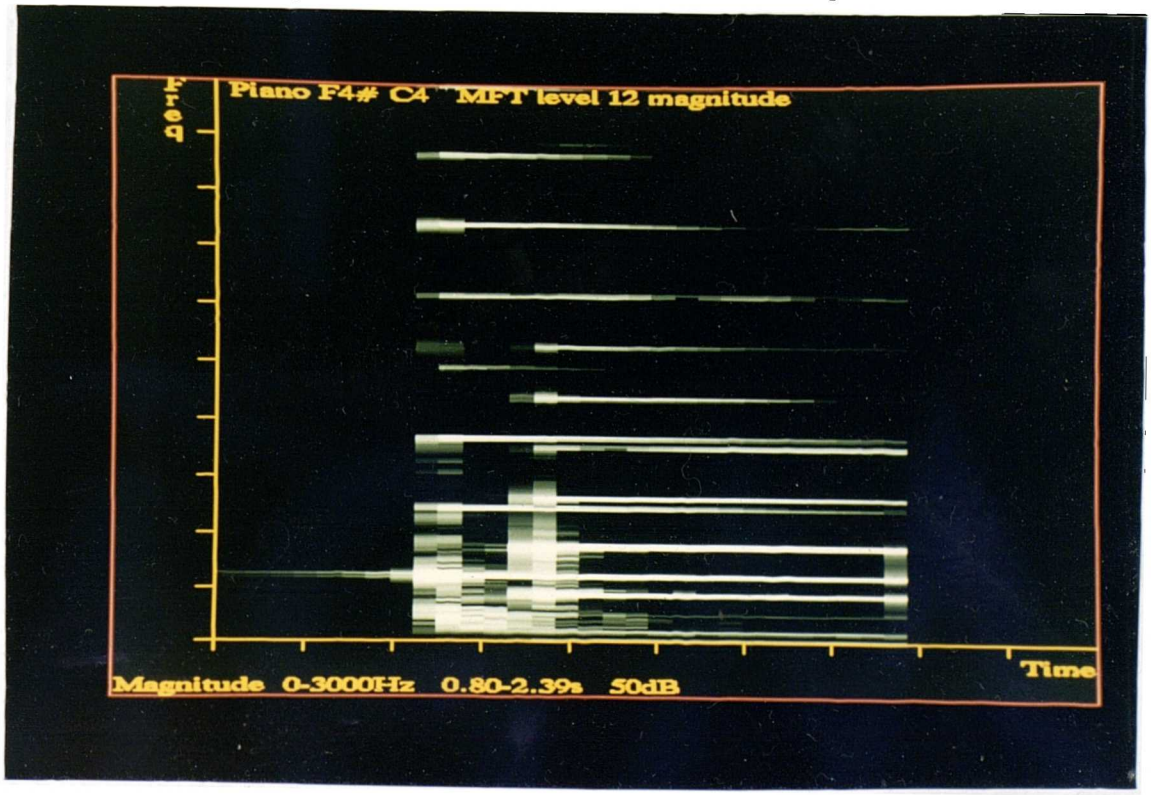

Figure 7.6: Two Piano Notes: MFT level 12, magnitude

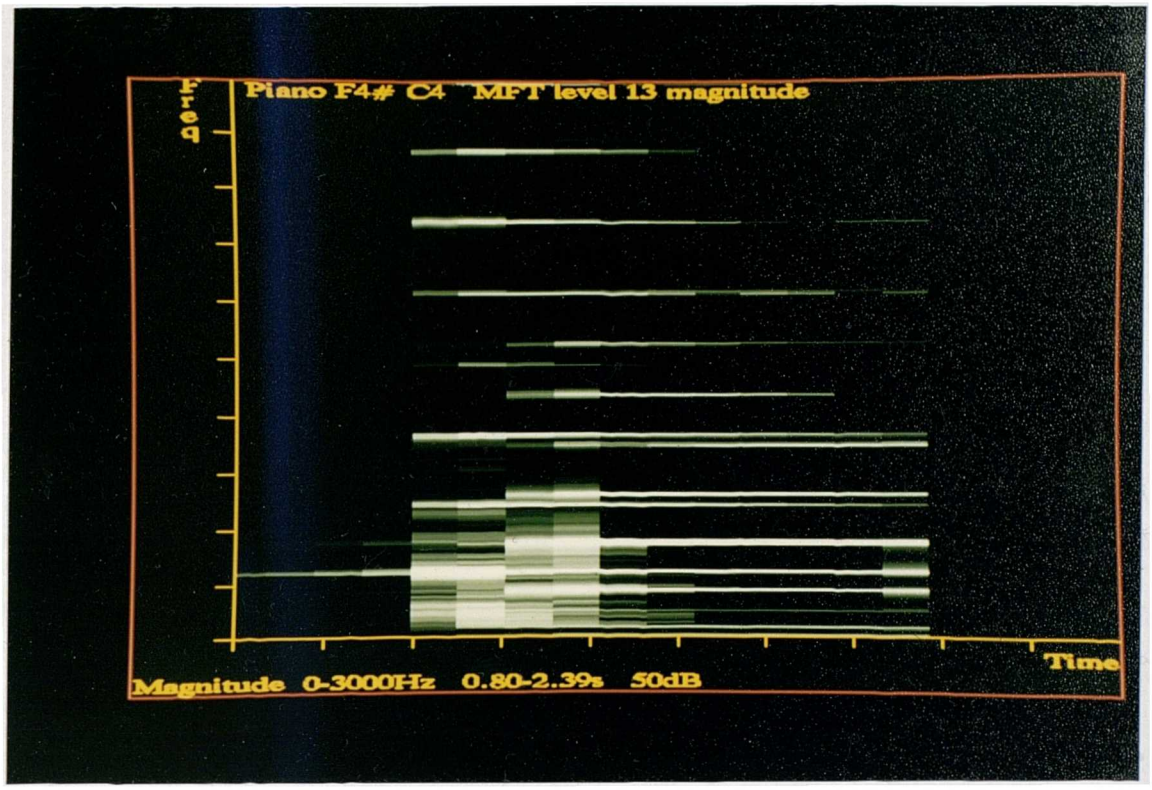

Figure 7.7: Two Piano Notes: MFT level 13, magnitude

in levels 11 to 13 by the 'leaking' of energy from the first partial of the first note across time into the coefficients before its onset. If the primary bin of this partial has a frequency index k, note that the leakage is relatively larger in the coefficients with frequency indices $k-1$ and $k+1$ than k. This is due to a discontinuity in the frequency variation of phase at the primary bin in the time frames close to the note's onset. This behaviour is useful since the leakage does not conform to the partial model defined in Chapter 4 and so can be discriminated from an actual partial.

Viewing the plots as a set, the harmonic nature of the signal can be seen at all levels. The lower levels do not have sufficient frequency resolution to separate adjacent partials. The beating which this causes is strongly evident in several places at the lowest level. Note that the 'difference' frequency of this beating covers a wide range: the fifth harmonic of the first note exhibits, on Level 9, cancellation points 240ms apart but only 20ms apart for the second and seventh. Level 11 just manages to resolve the two closest partials, the third harmonic of the first note (1100Hz) from

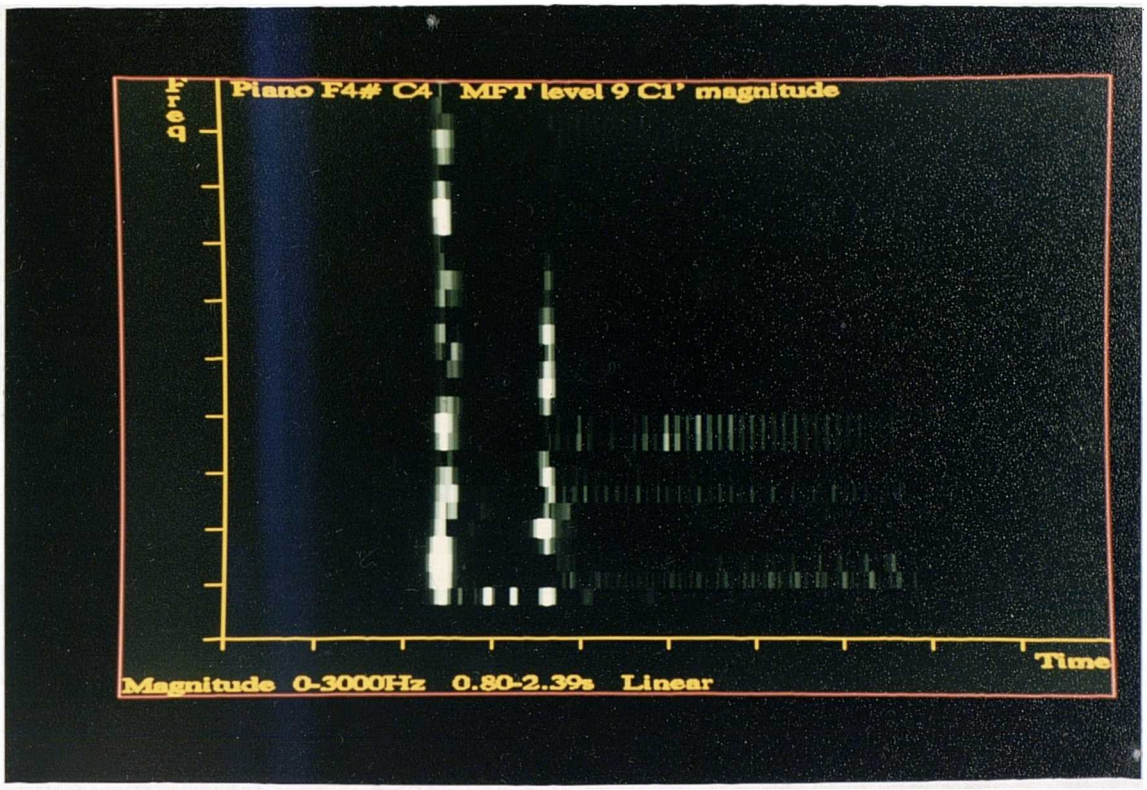

Figure 7.8: Two Piano Notes: $c'_{1ik}(9)$, magnitude

the fourth harmonic (1048Hz) of the second. Moving up in scale, the partials are better separated and resolved with respect to frequency, with a corresponding decrease in the clarity of the partial onsets. At level 13, which has the lowest temporal resolution, it is difficult to determine visually the relative onset times of the lower partials, though this is in part due to the compression of the top end of the amplitude scale in these plots.

7.3.2 Feature Detection

The transcription process proper starts from the raw MFT coefficients and generates the three measures $c_{0ik}(n)$, $c_{1ik}(n)$ and $c'_{1ik}(n)$, described in Chapter 5 for each coefficient $\hat{x}_{ik}$ on level n. Normally, these intermediate values are calculated on demand by the onset detection algorithms: no permanent record is made. For this case $c'_{1ik}(n)$ has been plotted in figures 7.8 to 7.10, since observation of this measure aids understanding of the onset detection process.

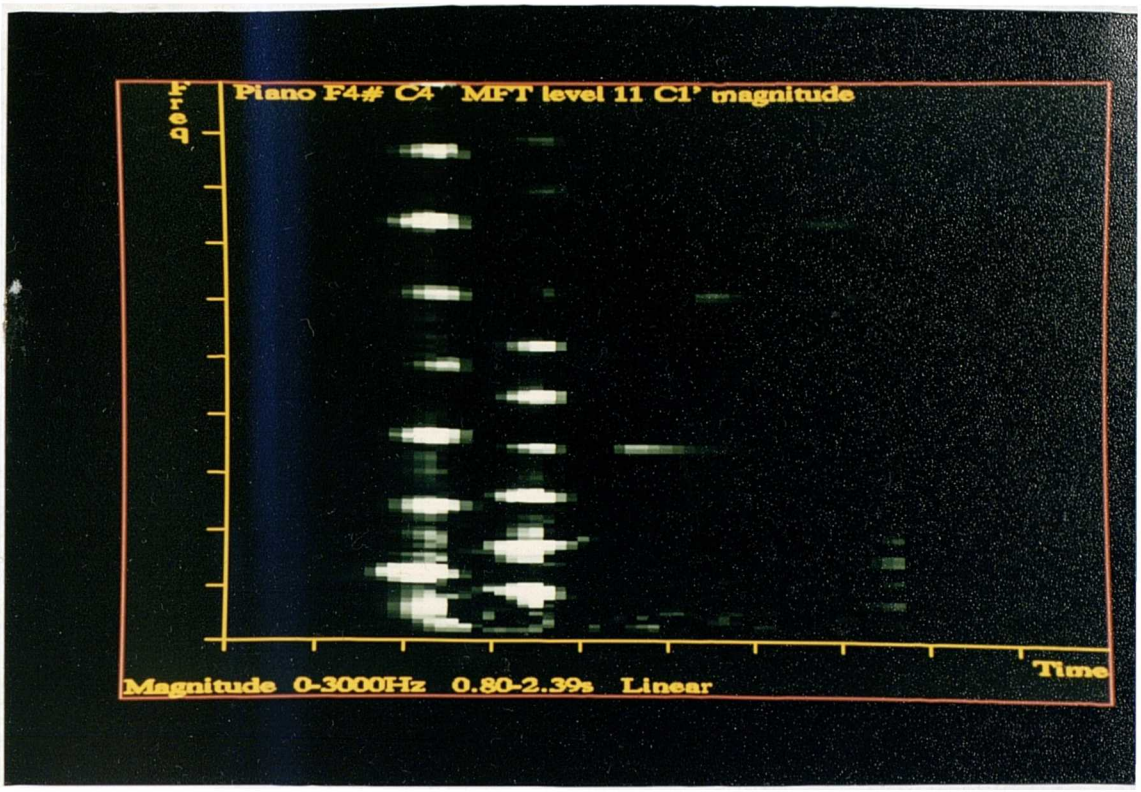

Figure 7.9: Two Piano Notes: $c'_{1ik}(11)$, magnitude

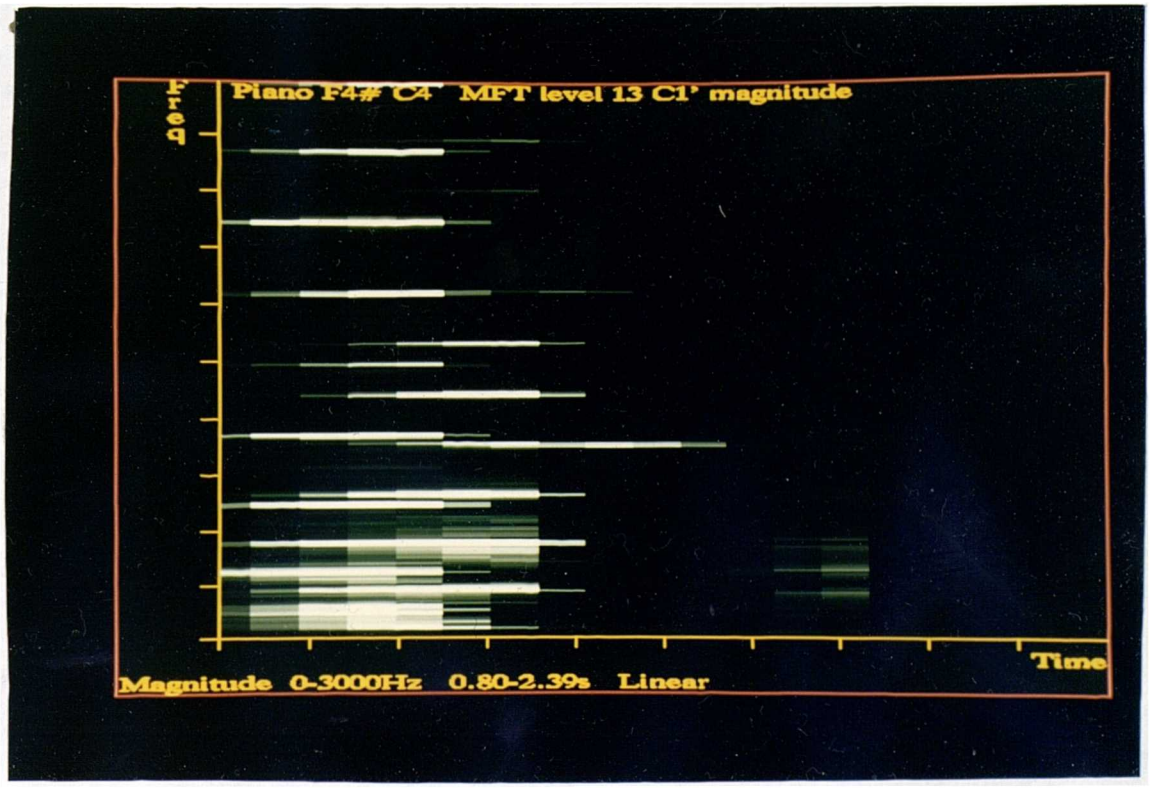

Figure 7.10: Two Piano Notes: $c'_{1ik}(13)$, magnitude

The resulting plots show the effect of the onset detection calculations on the original coefficient magnitudes. The steady-state portions of the signal have been largely suppressed, while peaks are centred around each partial onset. Comparing the levels, it can be seen, as would be expected, that the detector only works effectively at the resolutions at which the onsets are clearly resolved. On level 9 the detector responds strongly to the beating between the third harmonic of the first note and the fourth of the second note. Random responses caused by the low frequency noise are present at the lower levels. Although the highest level appears very cluttered, a large response has been produced for all of the partial onsets, admittedly with low temporal resolution.

The next stage of processing is to search for peaks in the response of the onset detection to generate a set of onset events for each level. These event sets are plotted in figures 7.11 to 7.15. The event markers are overlaid on the original coefficient magnitudes to show the alignment between them and the signal transform. Each event is represented by a green ellipse centered on its estimated time-frequency position. The relative lengths of the ellipse diameters serve to indicate the level of uncertainty associated with the event's time and frequency estimates. At this stage these uncertainties are based purely on the geometry of the level coefficients and so are of longer duration and less bandwidth with increasing level. Each event is annotated with its frequency in Hz.

The events generated on each level are a fairly good match with the partial onsets. In accordance with the requirements of the multiresolution event combining process, most of the errors are inclusion errors: events generated at points where there is no actual onset. Some events are excluded, however, due to lack of resolution in either domain causing the onset to merge with some other nearby feature. A notable example of this behaviour is the failure to detect the first partial of the second note at levels 9 and 10, as it is obscured by the nearby (in frequency) first partial of the first note. There are instances where the detector seems to be working better than was initially expected. The onset of partial four of the second note is closer in frequency to the first note's third partial than in the previous case; yet it is correctly detected at level 10 and above. This may be due to the use of both

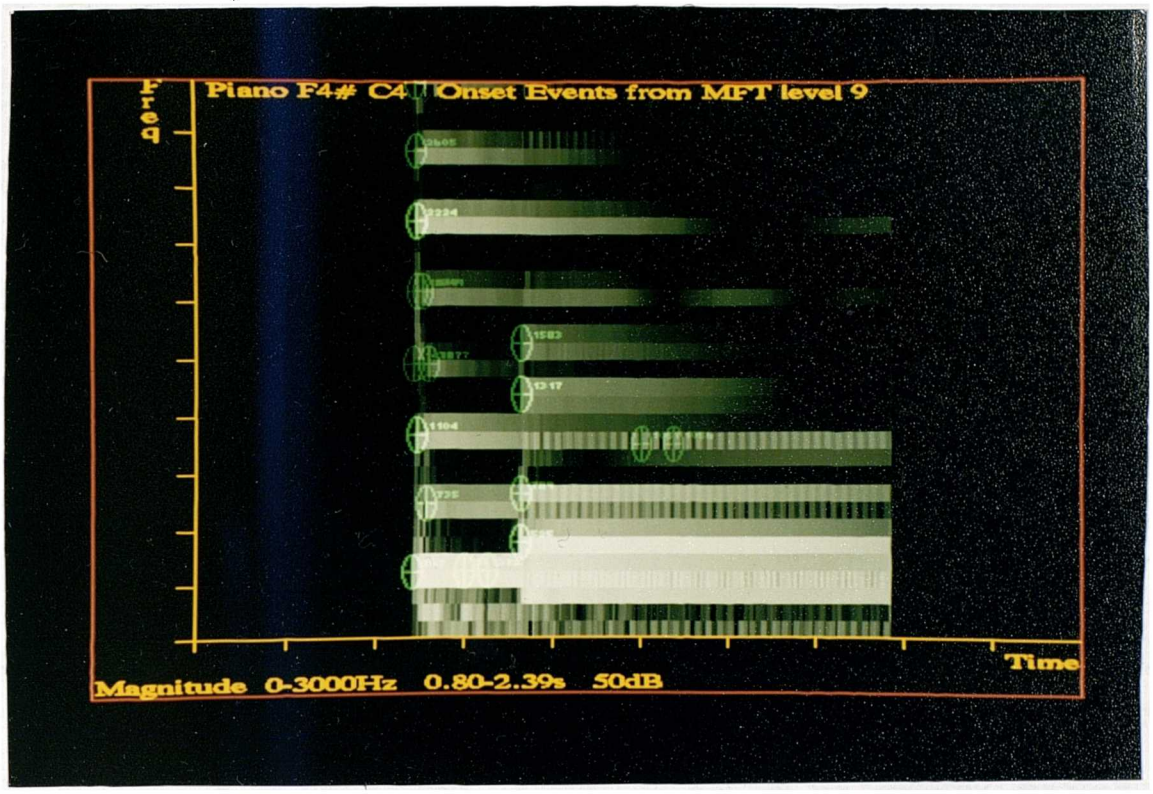

Figure 7.11: Two Piano Notes: onset events, level 9

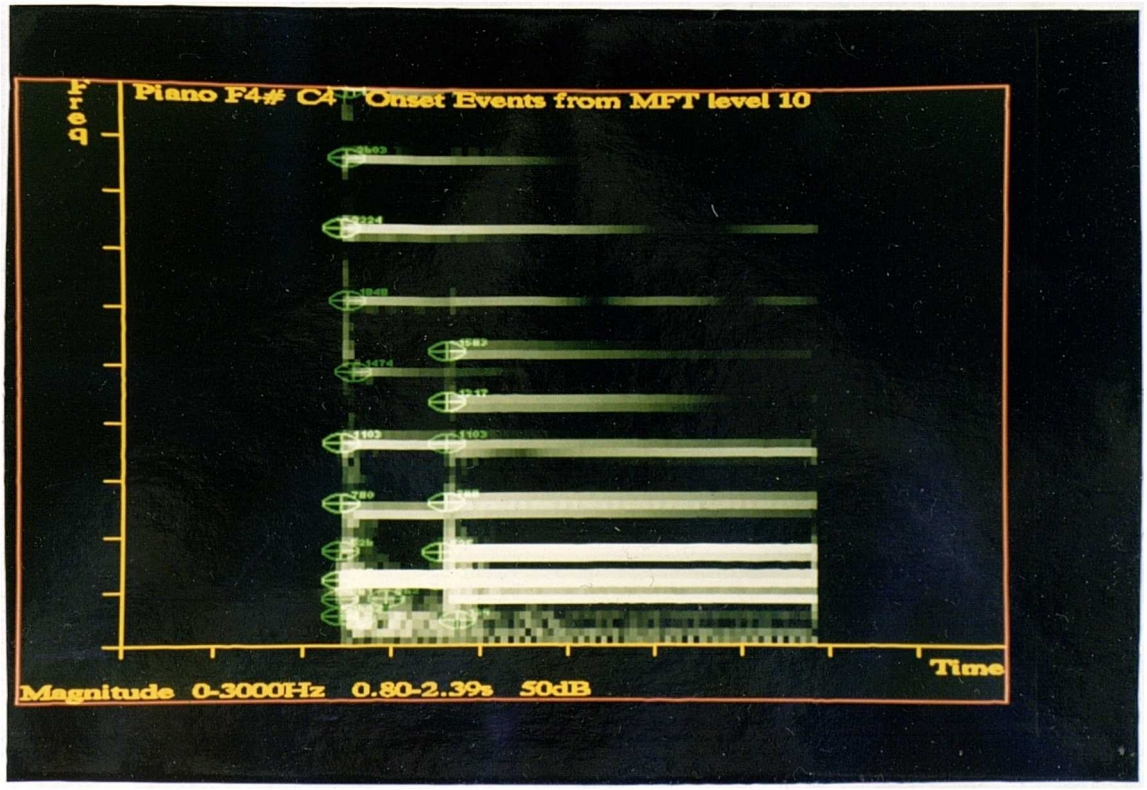

Figure 7.12: Two Piano Notes: onset events, level 10

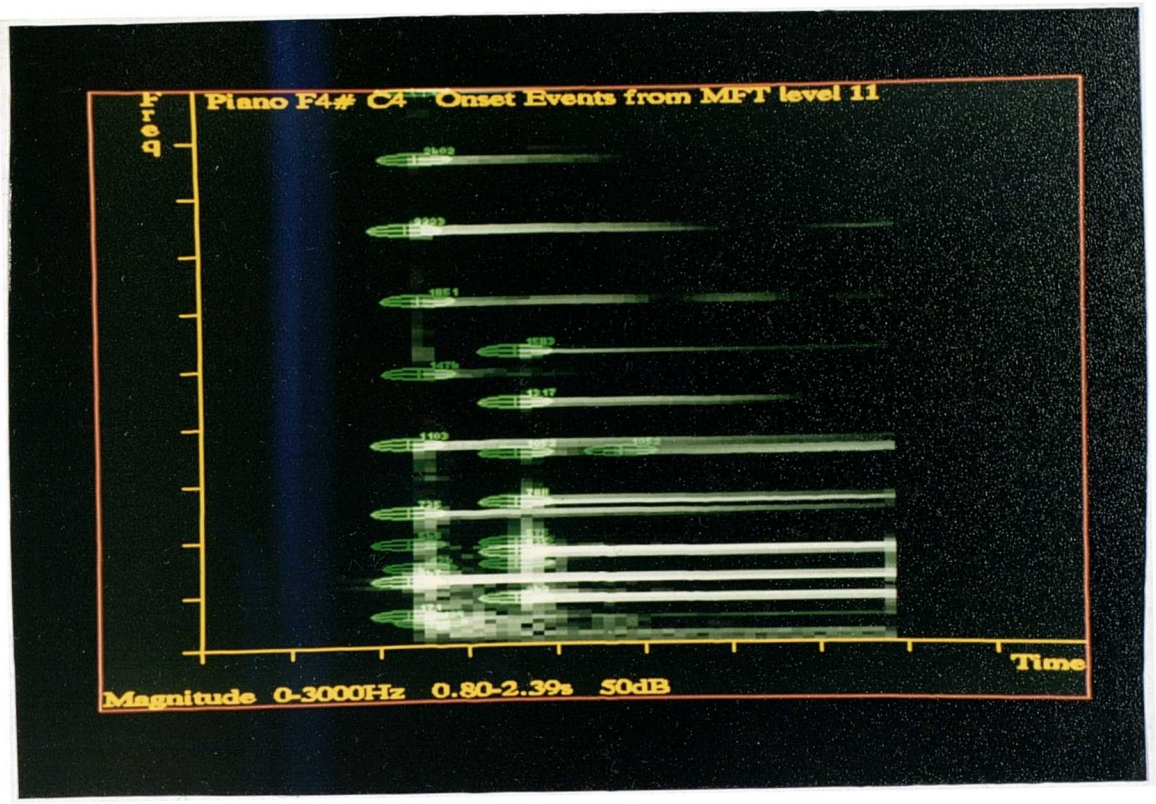

Figure 7.13: Two Piano Notes: onset events, level 11

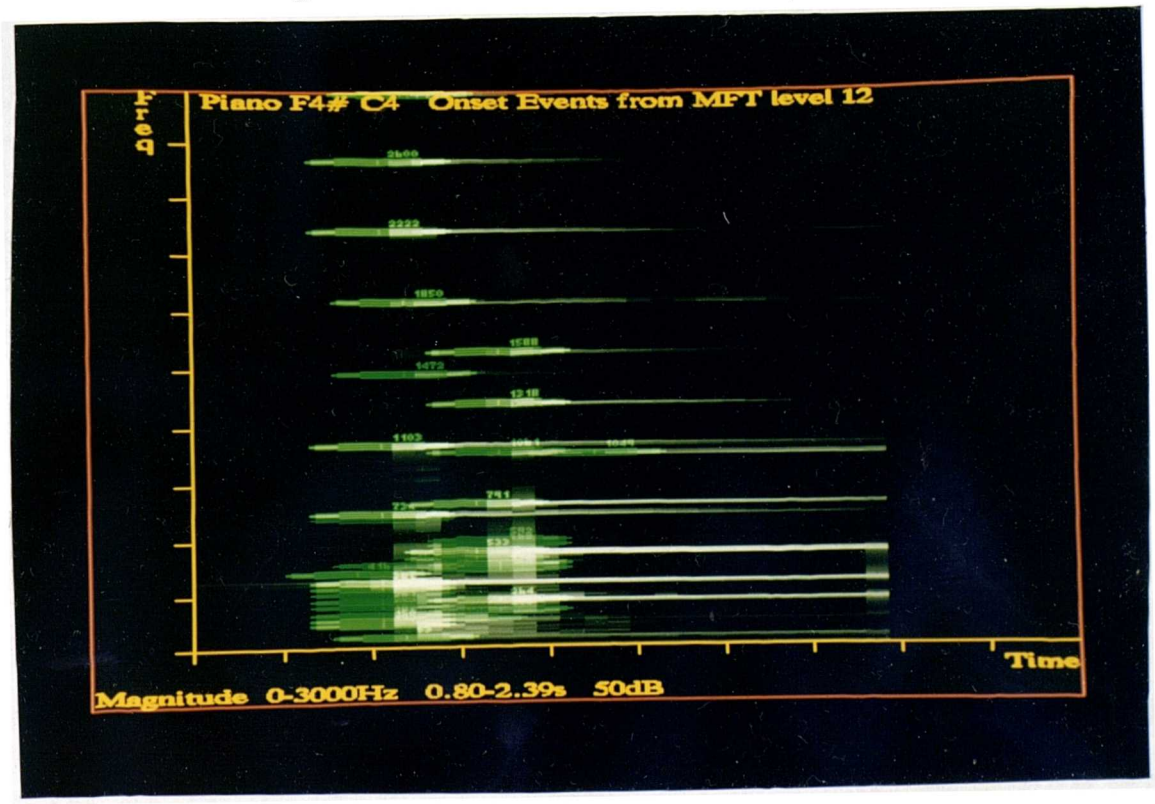

Figure 7.14: Two Piano Notes: onset events, level 12

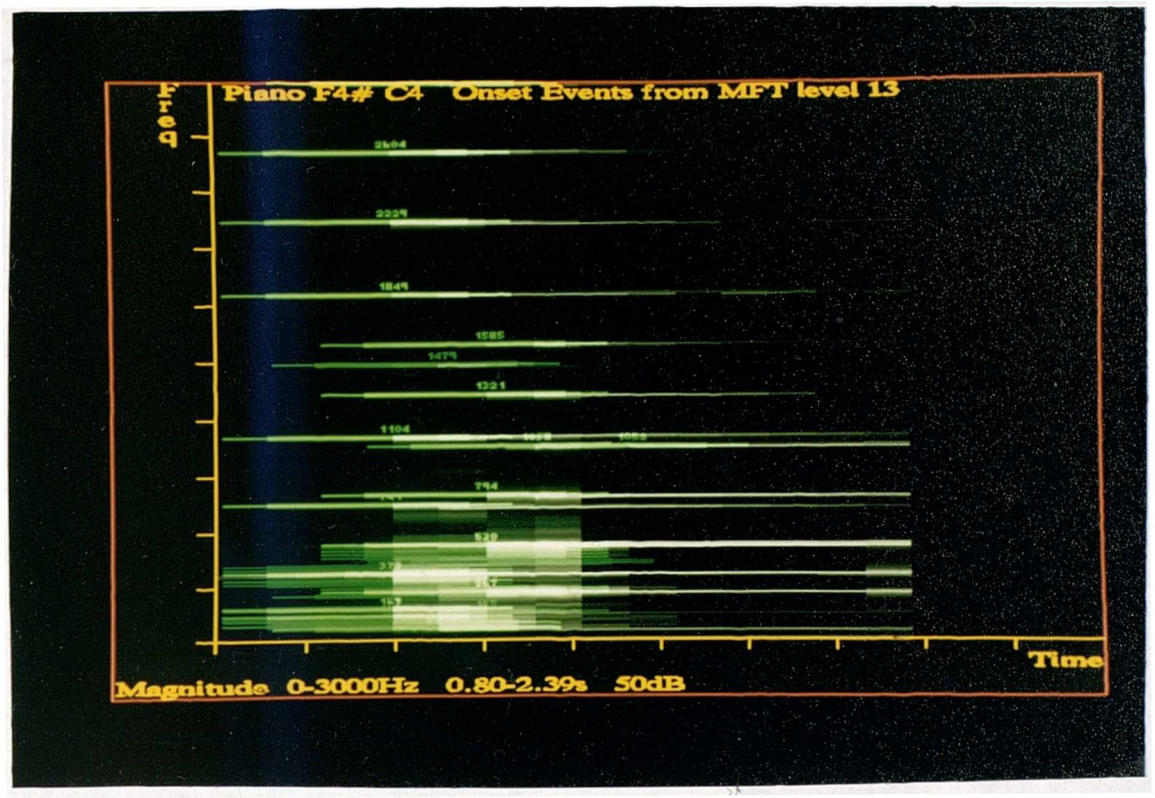

Figure 7.15: Two Piano Notes: onset events, level 13

magnitude and phase in the detection algorithms. Note that none of the partials above the sixth of the second note has been detected; this is due to their low magnitude.

The inclusion errors can be attributed to the low frequency onset noise, cancellation due to interference between poorly resolved partials and to multiple events being generated for a single onset. It is thought that more work on the peak detection algorithm could eliminate these errors. The next stages of processing, however, successfully remove these false alarms.

7.3.3 Multiresolution Processing

All processing up to this stage has been carried out on each level separately. The large amount of data present on each level has been reduced to much smaller sets of onset events, allowing the multiresolution processing to proceed relatively efficiently. Two sets of results, figures 7.16 and 7.17, are presented for this process to demonstrate the effect of the transient onset extension to

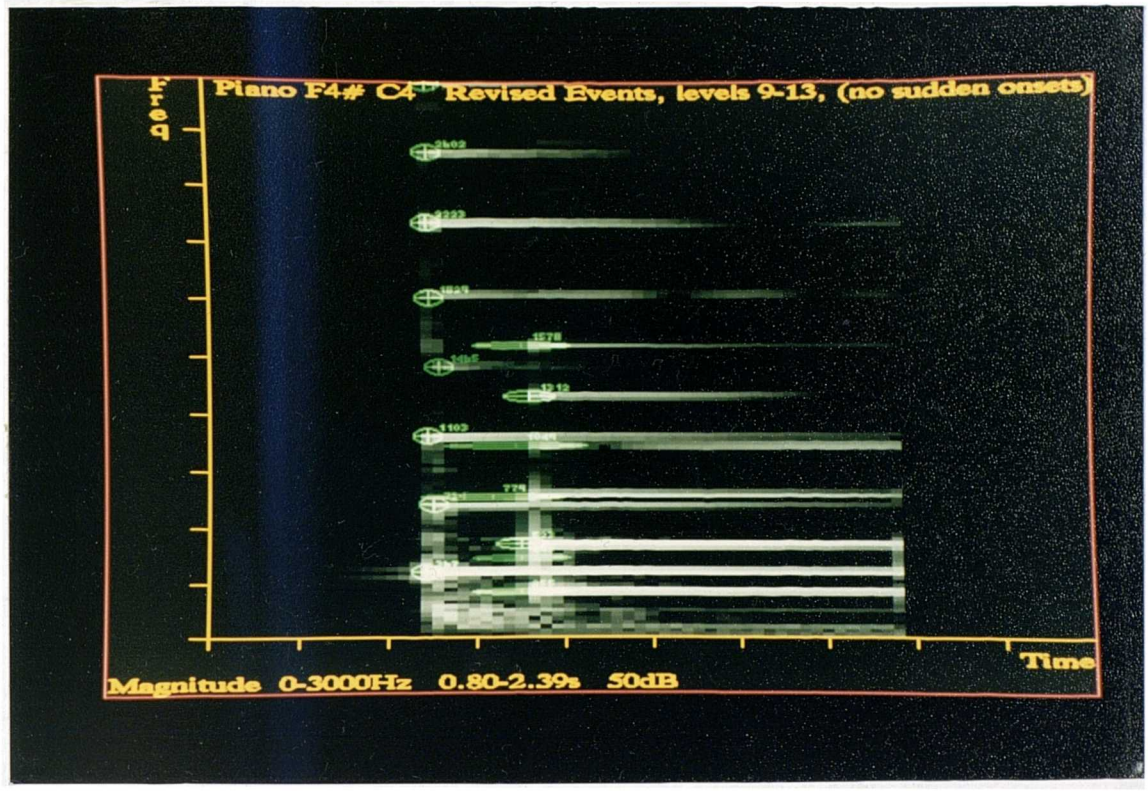

Figure 7.16: Two Piano Notes: Combined onset events, no transient detector

the onset detection described in sections 5.2.2 and 6.3.1.

Figure 7.16 shows the result of the onset combination without the transient onset phase detection. The algorithm, described in Chapter 5, has linked together nearby events across the levels to give a single set with decreased uncertainties. The varying shapes of ellipses can now be used to compare the relative uncertainties, in both domains, of the combined events. The effect of the determination of suitable resolution within the combination algorithm can be seen on partials three and four of the second note. These onsets have relatively large uncertainties associated with their onset time estimates, due to nearby features making the lower levels of the MFT unsuitable for reliable detection. There is still one inclusion error in this set, which is just below the onset of partial two of the second note.

The effect of the transient onset detector operating on the frequency variation of phase of the onsets can be seen by comparing figures 7.16 and 7.17. These plots are similar in all respects except

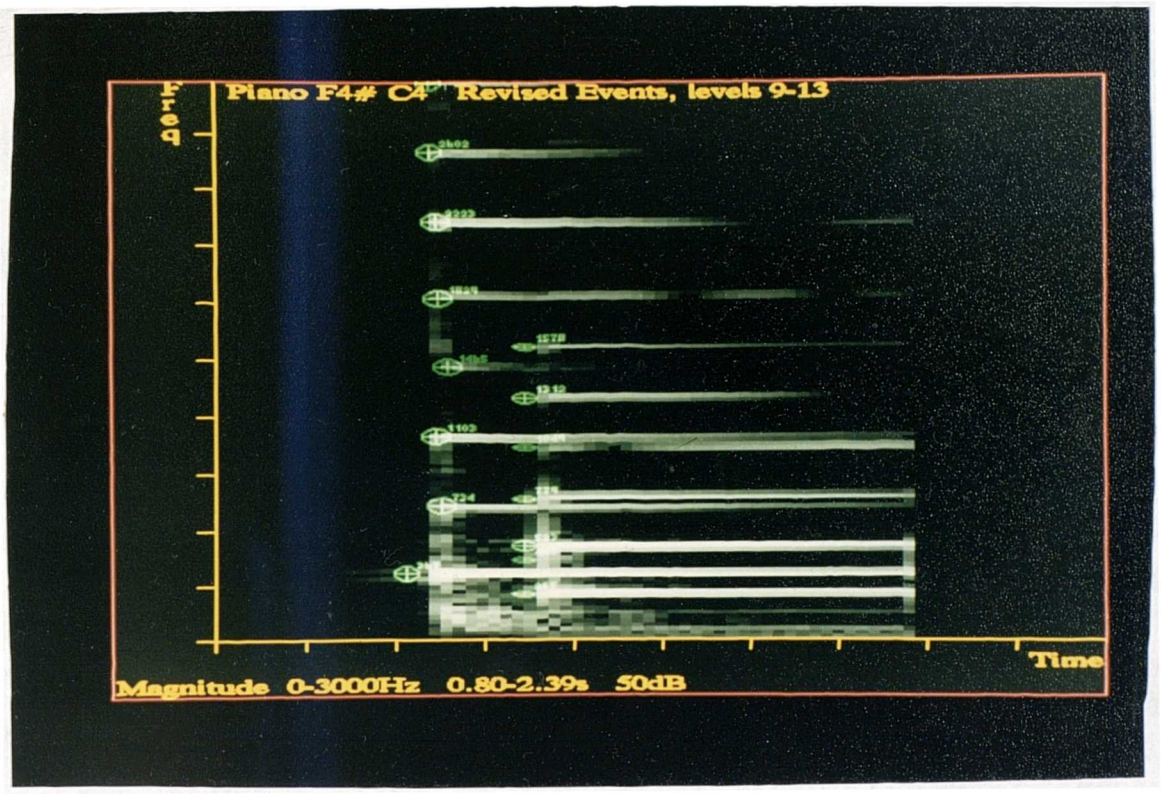

Figure 7.17: Two Piano Notes: Combined onset events, with transient detector

the addition of the transient onset detector. Note the way that the onset events which previously had relatively large temporal uncertainties due to nearby features, such as the onset of the fourth harmonic of the second note (1049 Hz.), have had that uncertainty decreased by this process. One disadvantage of this is an increase in the susceptibility of the detector to 'ghosting' effects caused by the temporal sidelobes of the FPSSs. This is exhibited by the moving forward in time of the onset event for the first harmonic of the first note.

The determination of the partial onset events proceeds simultaneously with the generation and tracking of partials, as shown in Figure 7.18. The partial magnitude is not included in this figure, so that the positions of the partials may be compared with the level coefficients (from level 11).

Figure 7.19 is a more representative illustration of the detected partials. In this plot, the intensity of the line represents the magnitude of the partial, while its colour shows the level which has been selected for tracking. Note the way in which the start of the second note causes the tracking levels

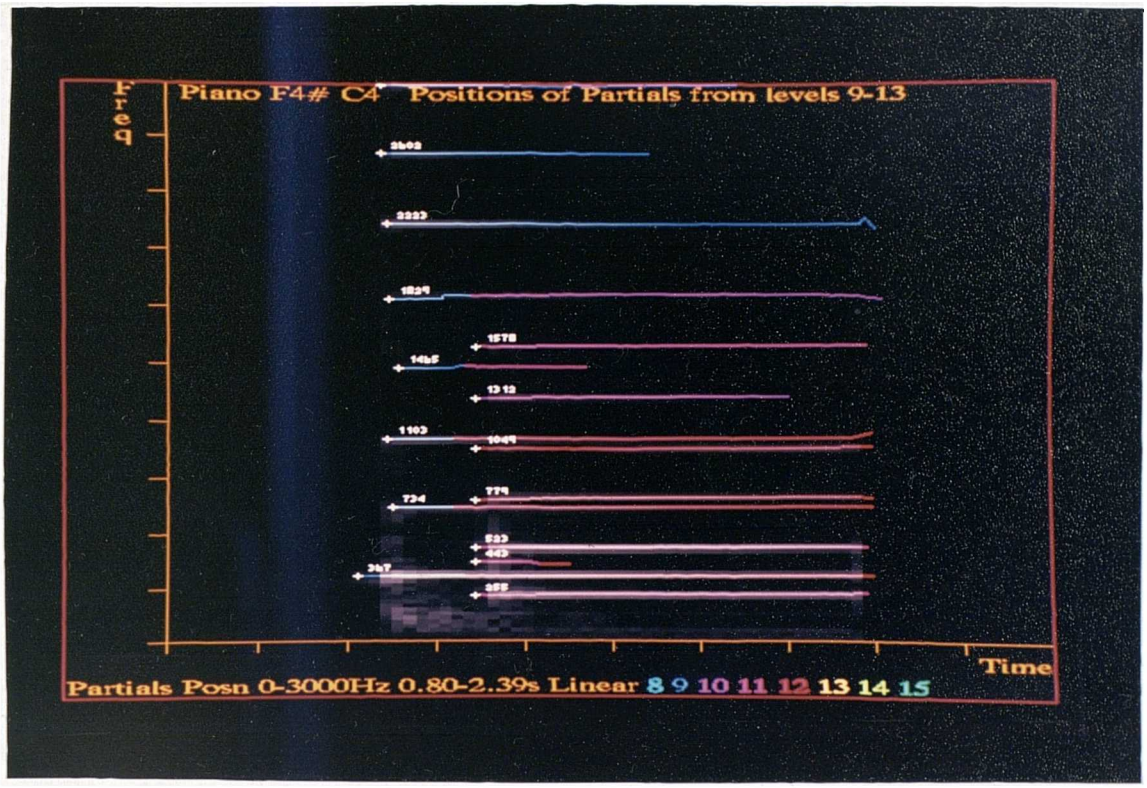

Figure 7.18: Two Piano Notes: Partial Positions

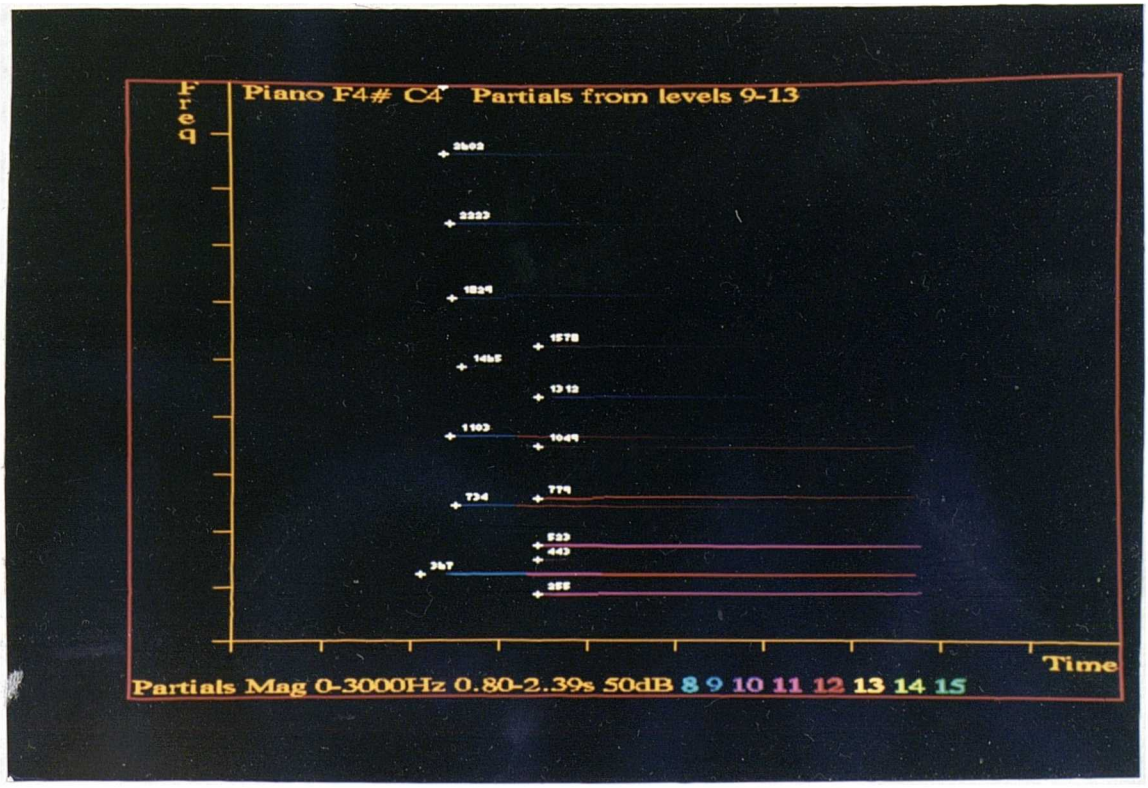

Figure 7.19: Two Piano Notes: Partials

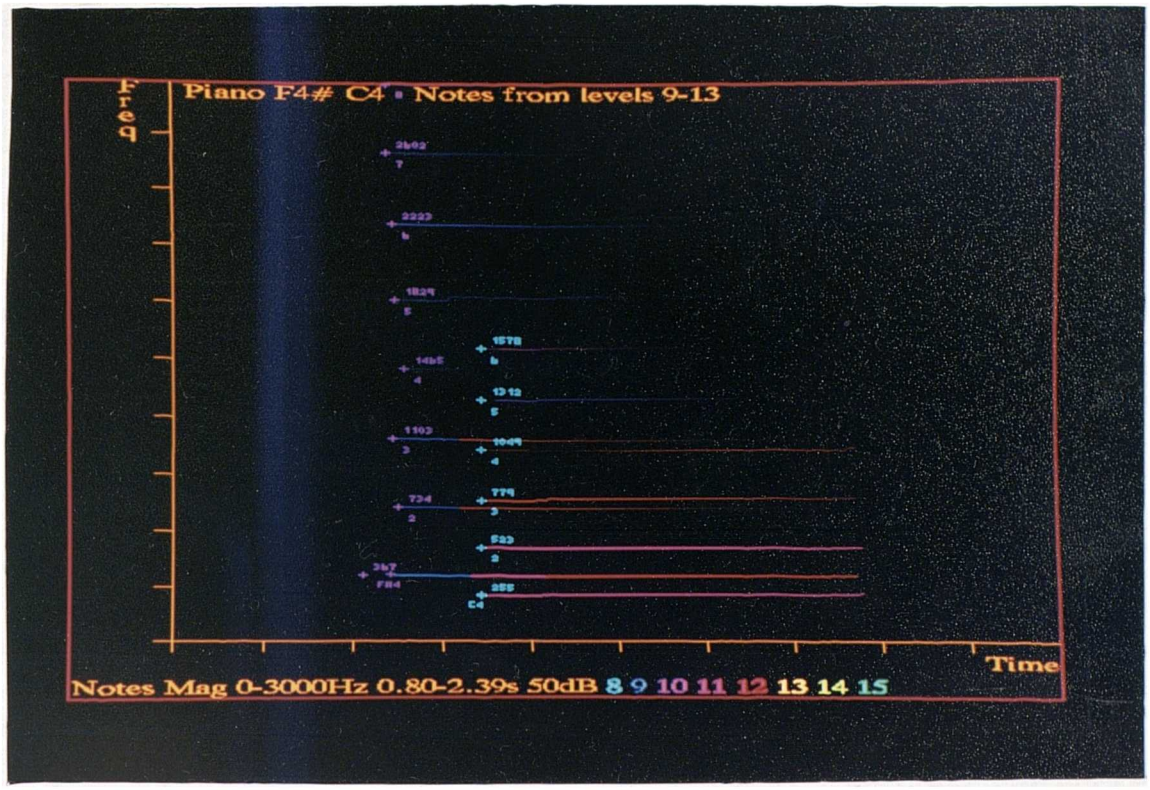

Figure 7.20: Two Piano Notes: Notes

of the first note to increase. The single onset inclusion error generates an incorrect partial, but the tracking process terminates it after a relatively short period.

The final step in the analysis is to associate groups of partials with notes. This is relatively easy in this case, as is shown in Figure 7.20. Here, the colour of the text at the onset of each partial indicates its note grouping. The note name is indicated on the fundamental and the harmonics display their number. The single erroneous partial has been rejected, since it could not be associated with any other partials.

The analysis of the two piano notes resulted in both notes being identified, with partial tracks generated for the lower eight harmonics of the first note and the lower six harmonics of the second note. The estimated and expected partial frequencies are given in table 7.1. The estimated frequencies assume that the instrument was tuned to A440 and that the partials are an exact harmonic sequence. The latter assumption is not quite true; it is recognised that piano partials are sharpened

Note	Partial	Expected (Hz.)	Estimated (Hz.)
F4♯	1	370	367
	2	740	734
	3	1110	1102
	4	1480	1465
	5	1850	1829
	6	2220	2223
	7	2590	2601
	8	2960	2985
C4	1	261	255
	2	522	523
	3	783	779
	4	1044	1048
	5	1305	1312
	6	1566	1576

Table 7.1: Expected and Estimated partial frequencies for two piano notes

at high frequencies [Bla65], this does, however, seem to be reflected in the results. The estimated pitches for the notes are 370 Hz for the F4♯ and 262 Hz for the C4; which are their expected values.

The estimated onset time for the first note is 1.18s and the second at 1.40s, within 20ms of the actual times.

For this example, as can be seen from Figure 7.9, that all onsets and partials are resolved at level 11 and so it should be possible to obtain all the note parameters from that single resolution. The power of using many resolutions even for this simple case lies in the event confirmation and reduction of uncertainty provided by the scale space processing. There are 25 events detected on level 11 (fig. 7.13), of which 11 are false alarms: after combination (fig. 7.17) this is reduced to 14 correct events and one false alarm. The frequency uncertainties are, on average, reduced by a factor of 2 while the temporal uncertainties are reduced by at least a factor of 4.

7.4 Bach Woodwind Trio

The second set of results shows the analysis of just over two bars (5 seconds) from the start of Trio I for two oboes and bassoon which forms part of the first Brandenburg Concerto by J.S. Bach.

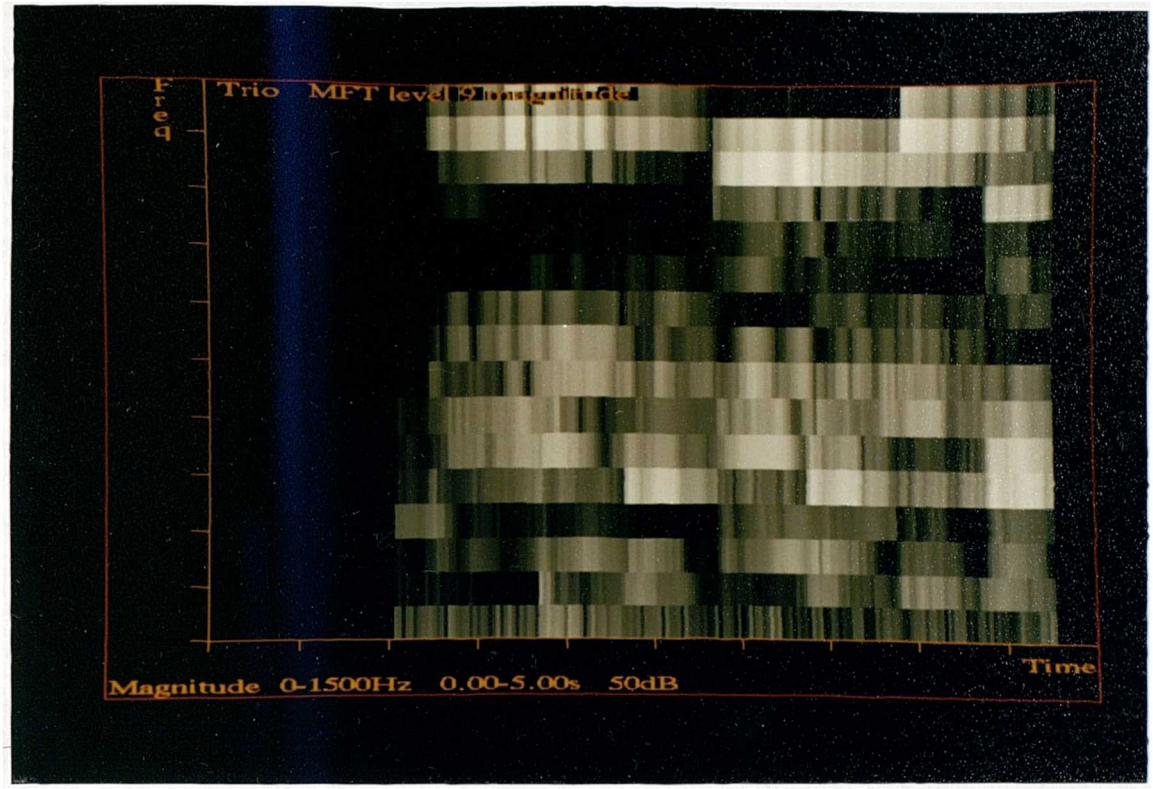

Figure 7.21: Bach Trio: level 9 magnitude

Figures 7.21, 7.22 and 7.23 show the coefficient magnitudes for levels 9, 11, and 13 for this signal. Clearly this signal is a great deal more complicated than the previous example. There are three instruments playing simultaneously and since they are playing fairly quietly there is little dynamic range. The plots appear much less 'clean' than than those of the two piano note example. The large number of simultaneous partials results in almost no detail being revealed at level 9 — only the higher levels, with increased frequency resolution, fully reveal the detail.

The results of applying the onset detection to these levels are shown in figures 7.24, 7.25 and 7.23. Unsurprisingly, few detections can be made from level 9 and those that have been are largely incorrect. The detectors, however, do behave well at level 11 and above, detecting all onsets where the available time or frequency resolution allows. Most of the false detections are caused either by low frequency noise, beating between coincident partials or clusters of events around a single onset. As before, this last class indicates that improvements could probably be made to the

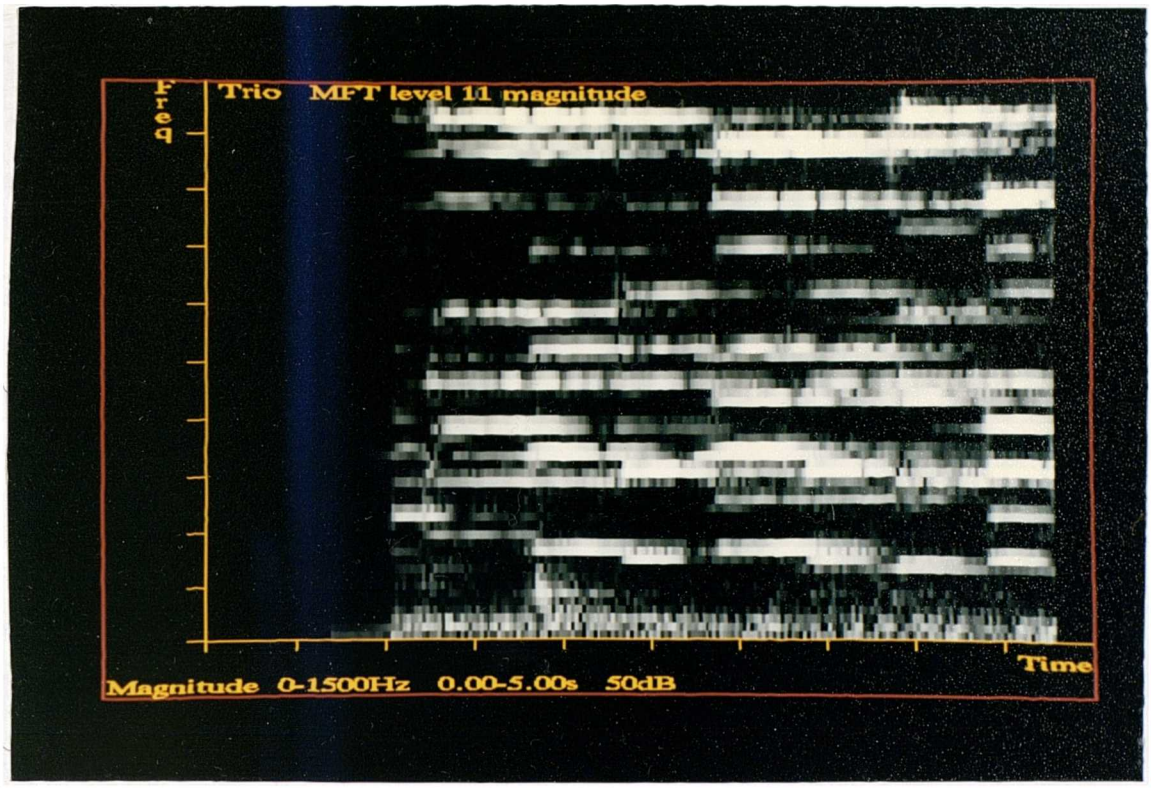

Figure 7.22: Bach Trio: level 11 magnitude

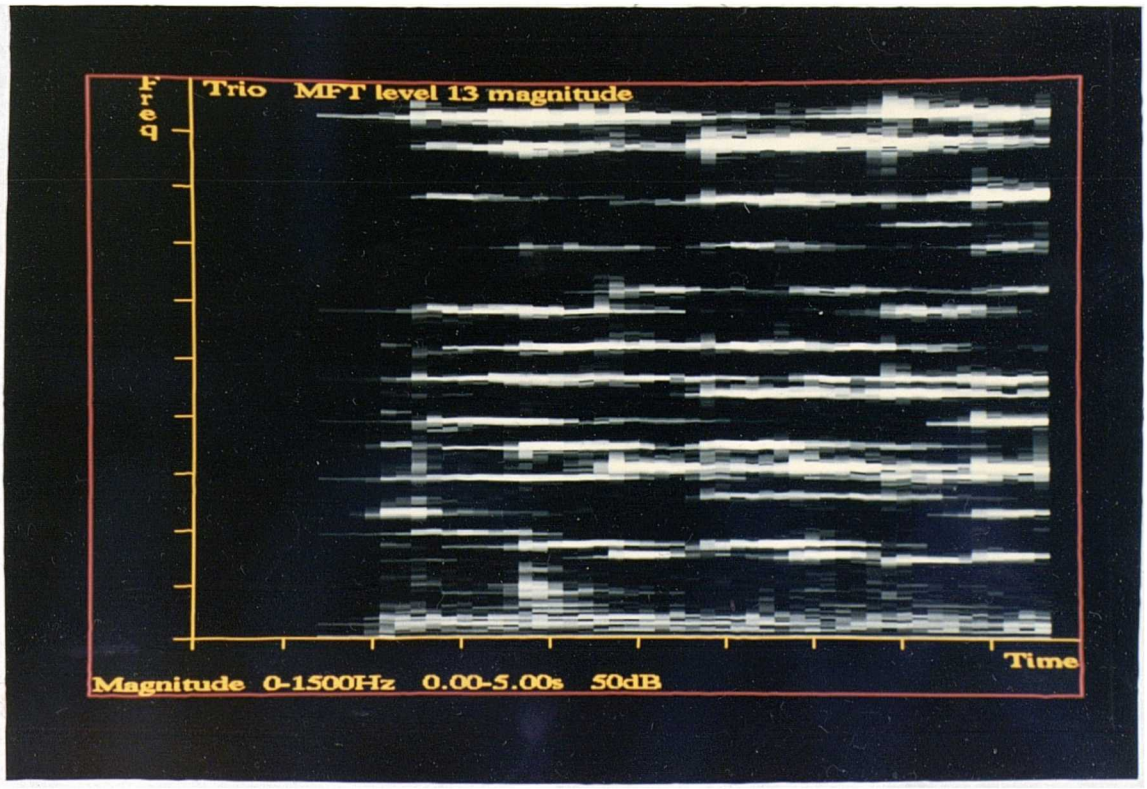

Figure 7.23: Bach Trio: level 13 magnitude

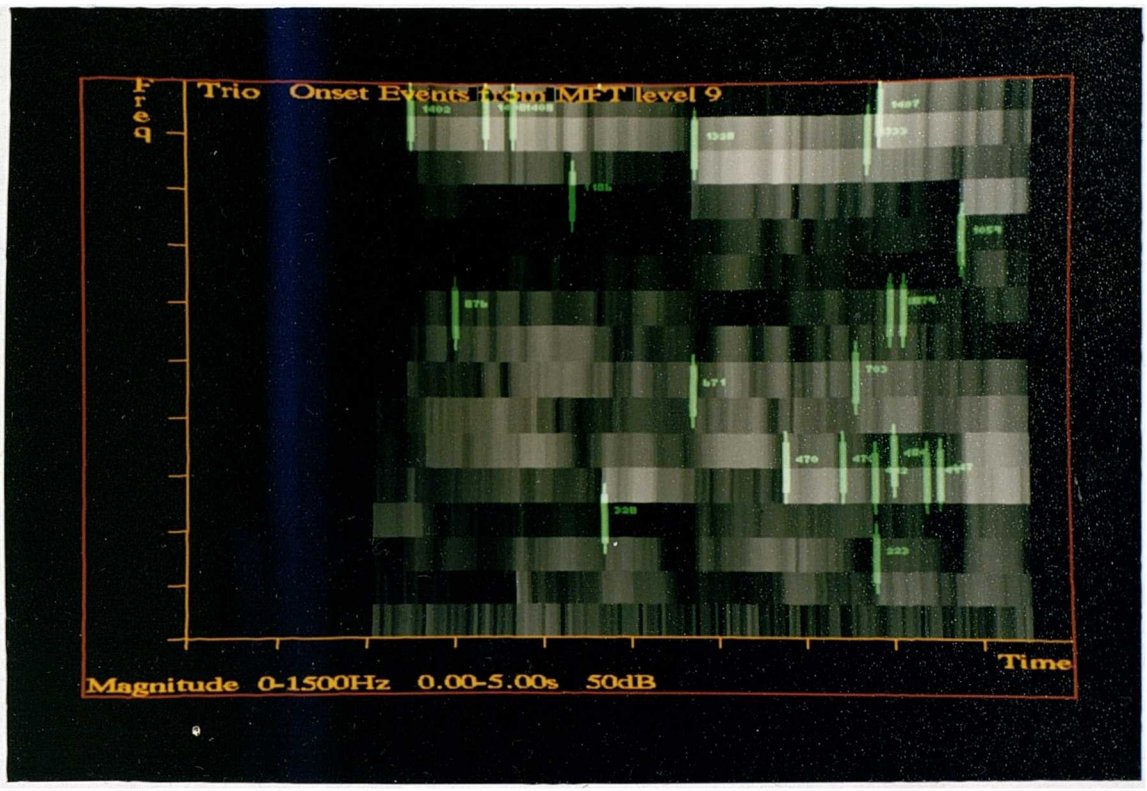

Figure 7.24: Bach Trio: onset events, level 9

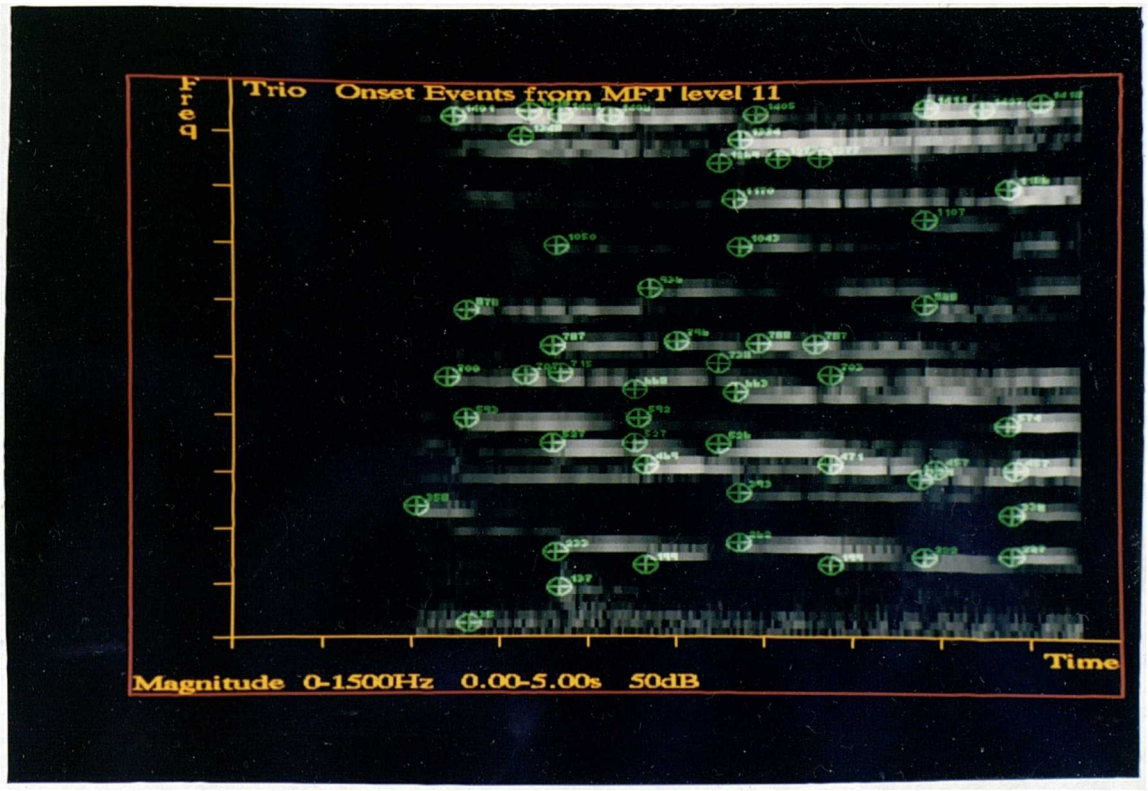

Figure 7.25: Bach Trio: onset events, level 11

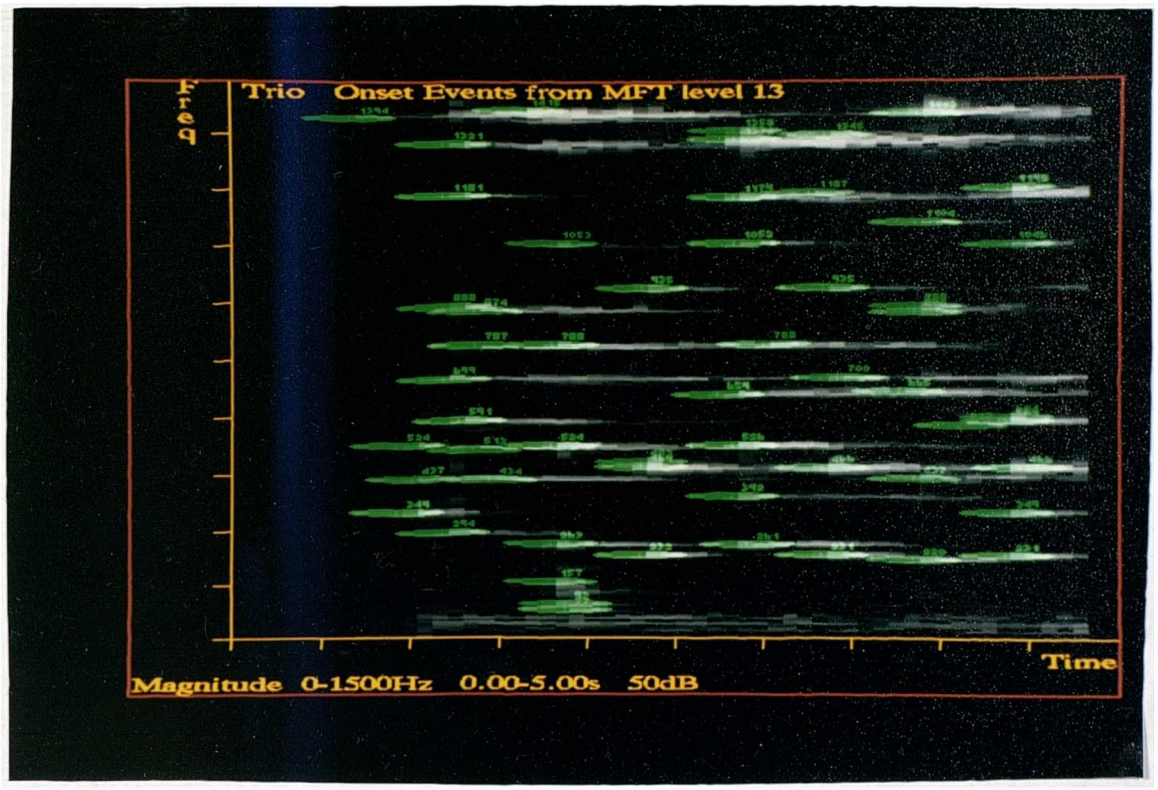

Figure 7.26: Bach Trio: onset events, level 13

onset detection algorithm.

The results of the multiresolution processing are shown in figures 7.27 and 7.28 which show the partial tracking levels and note allocation respectively. The notes of the original piece are given in table 7.2 while those detected by the analysis are shown in table 7.3. Not all the notes shown in the plot are included in the table, a threshold of 0.1 having been applied to the harmonic correlation (eqn. 6.46) to remove those note hypotheses with very weak harmonic structure.

	Instrument		
Bar	Bassoon	Oboe 1	Oboe 2
1	D4	A4	F5
	C4		
	A3♯	A4♯	G5
2	C4	G4	E5
	A3♯		
	A3	A4	F5
3	A3♯	F4	D5

Table 7.2: Notes from Bach Trio

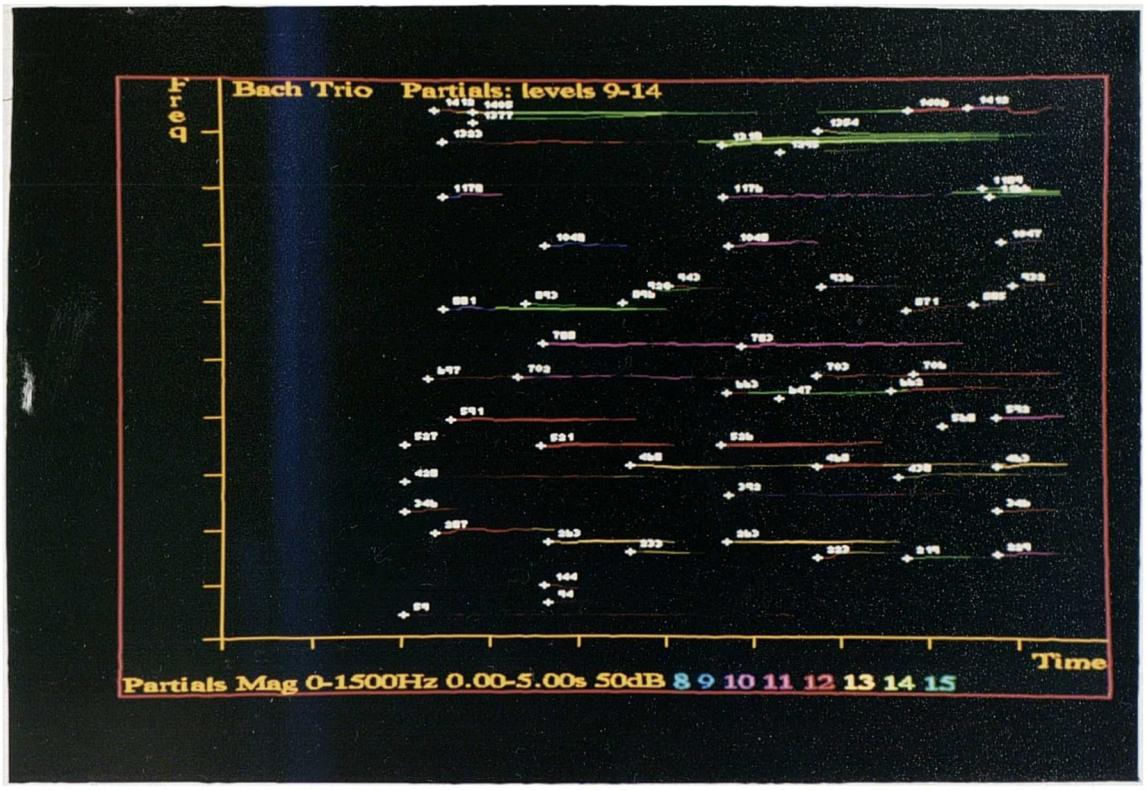

Figure 7.27: Bach Trio: partials with tracking levels

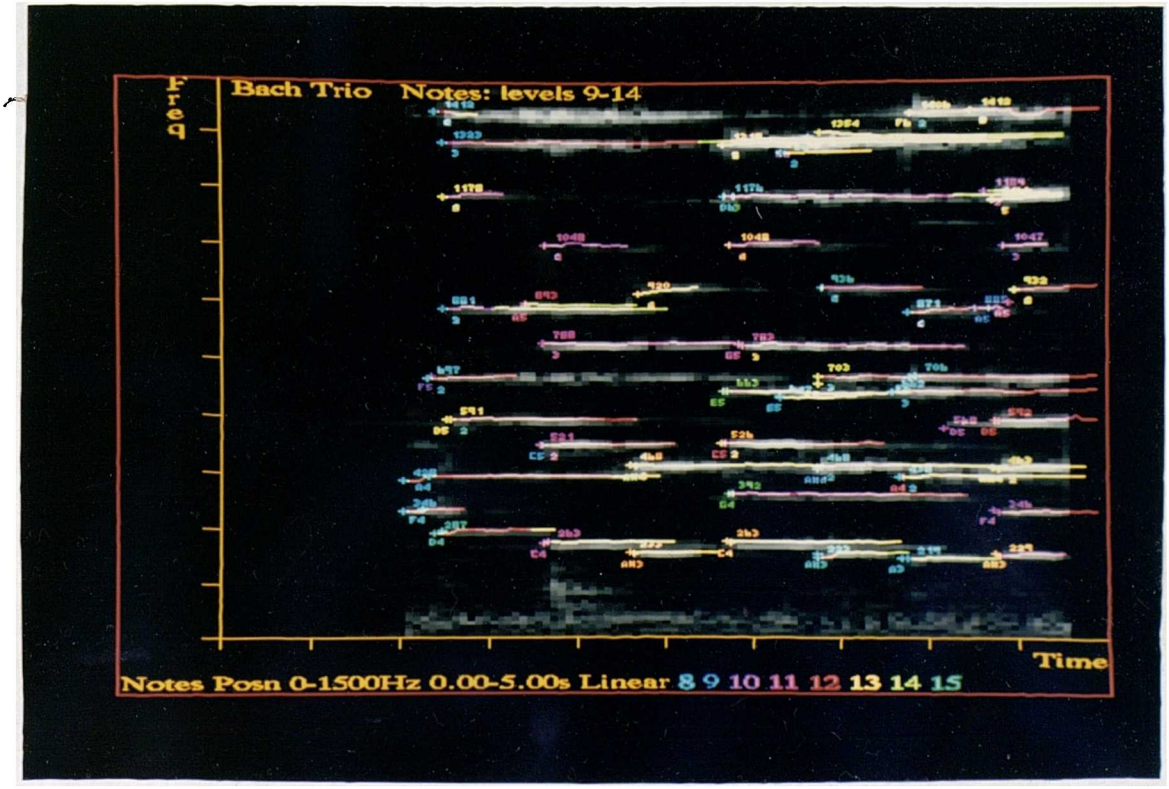

Figure 7.28: Bach Trio: note allocation and partial positions

Bar	Note Hypotheses				
1	D4	F4	A4	F5	
	C4	C5			
	A3♯	A4♯			
2	C4	G4	C5	E5	G5
	A3♯	A4♯			
	A3	A4	F5		
3	A3♯	F4	A4♯	D5	

Table 7.3: Notes from Bach Trio Analysis ($r_i > 0.1$)

Piano				Cello	Violin
D3♯	D4♯	D5♯	D6♯	D3♯	D5♯
A2♯	A3♯	A4♯	A5♯	A2♯	A4♯
G2	G3	G4	G5	G2	G4
G2♯	G3♯	G4♯	G5♯	G2♯	G4♯
C3	C4	C5	C6	C3	C5

Table 7.4: Notes from Schubert Trio

There are 17 notes in the original score, and 22 notes in the analysis. There is one note excluded, a G5 on the third beat of the first bar, which is masked by the C4 of the preceding beat being at exactly one third its frequency. Of the other (inclusion) errors, the F4 on the first beat seems to be caused by reverberation from the preceding movement, which was masked off for this analysis. All other inclusion errors are notes at one octave above correct notes. Without including factors such as the number and type of instruments playing into the analysis model such hypotheses are valid. A simple way to reduce these would be to increase the correlation threshold, but this would tend to remove correctly identified notes, such as those on the third beats of the first and second bars where two instruments *are* playing an octave apart.

7.5 Schubert Piano Trio

The final piece analysed is the first few notes of Schubert's piano trio in E♭ major (op. 100) for piano, cello and violin. The notes of the piece are given in Table 7.4. As can be seen, the instruments are playing in unison, which causes several partials to be coincident in frequency. This coincidence

D3♯	D4♯	A4♯	D5♯	G5	C6♯
A2♯	A3♯	F4	A4♯	F5	G5
G4	G5				
G4♯	G5♯				
A3♯	A4♯	A5♯			

Table 7.5: Notes from Schubert Trio Analysis ($r_i > 0.1$)

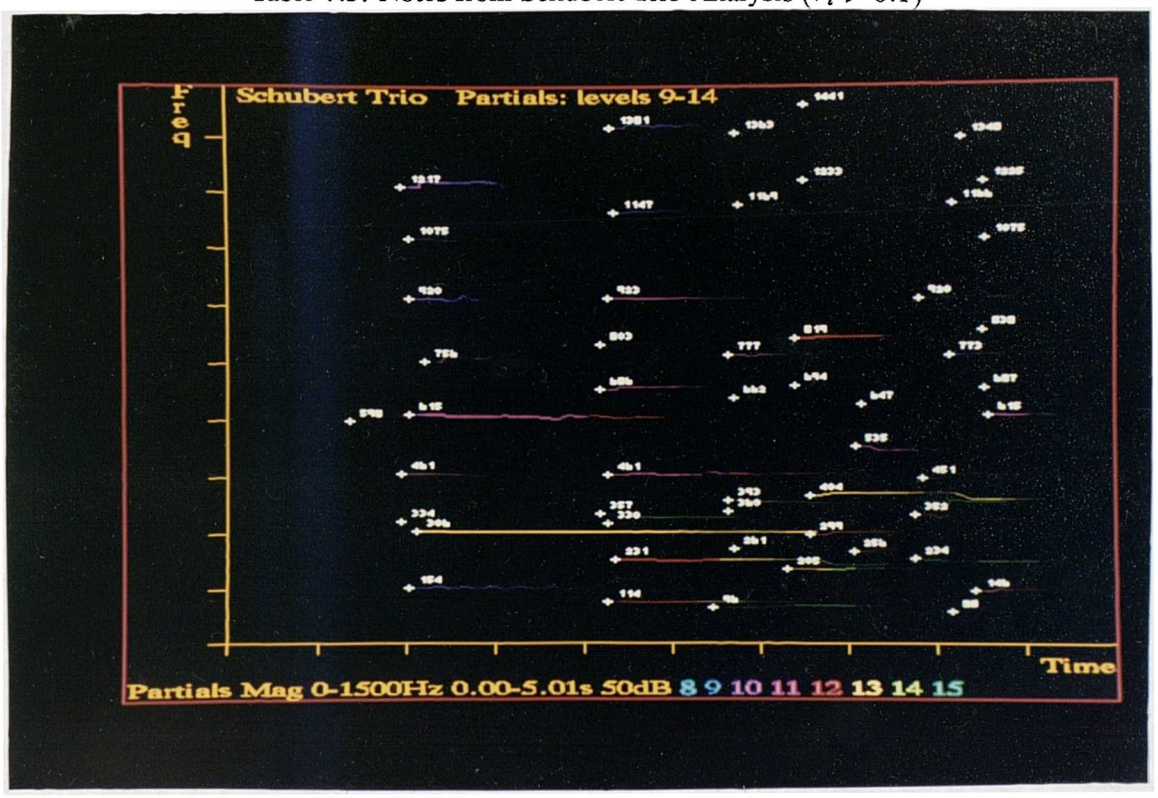

Figure 7.29: Schubert Trio: partials with tracking levels

destroys the phase coherence found in single partials and introduces irregular beating in the partial amplitudes. Additionally, the sample was recorded from an FM radio transmission onto cassette tape and consequently has a fairly high noise floor, estimated at around -45dB.

Figure 7.29 shows the partials detected the signal. The majority of partials appear to have been correctly identified and tracked, with only a few false detections. An increased amount of frequency jitter can be seen in the partial tracks, compared with the previous example, which can be attributed to the corruption of the partial phases. The partials which have not been identified have either their onsets masked by ongoing partials or are so affected by the lack of phase coherence, beating

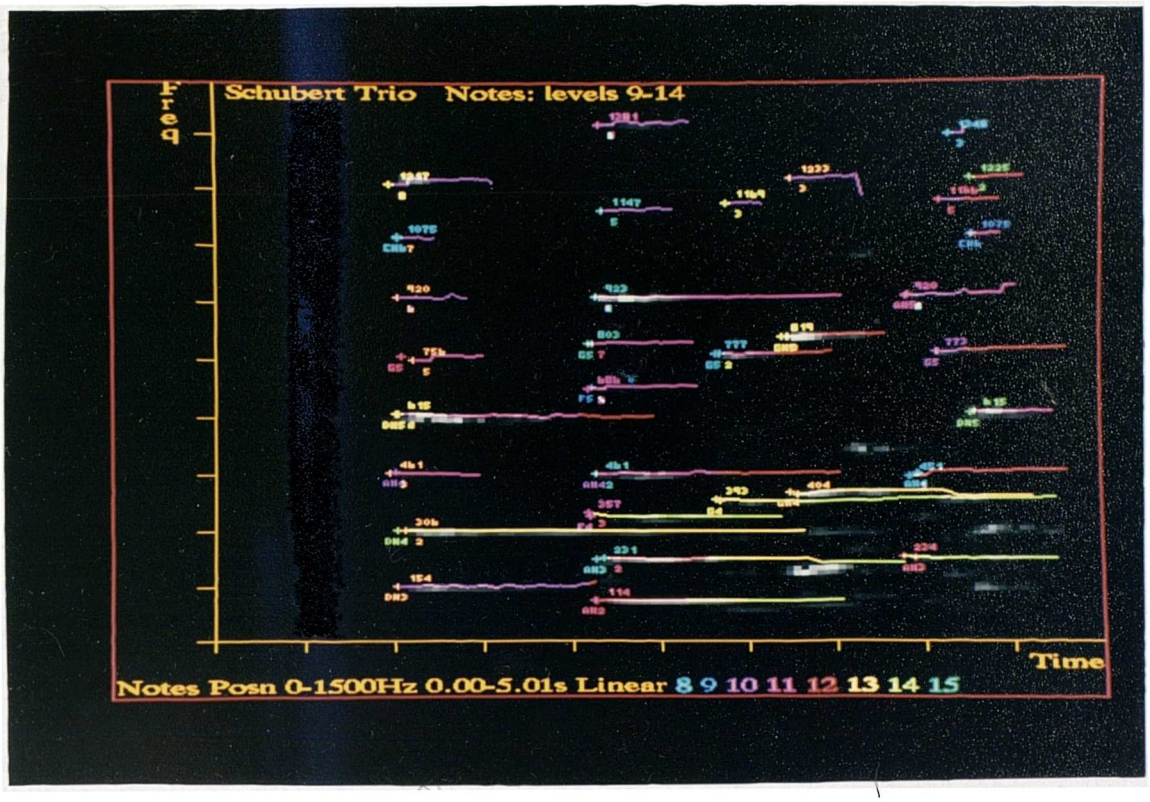

Figure 7.30: Schubert Trio; note allocation and partial positions

and noise to cause the onset detector to fail. The number of partials missed increases with time as the level of reverberation from the preceding notes interferes further with their already complex structure. Given that the detectors are designed to operate on phase coherent data, this part of the transcription process appears to work well with this demanding signal.

The note allocation result is shown in Figure 7.30 and Table 7.5. The allocation algorithm is heavily dependent on all partials being identified and is not sufficiently robust to deal with the many ambiguities introduced by several instruments playing in unison. The algorithm thus copes fairly predictably with the first two notes of the set; additional notes identified are all harmonically related to those present, while the failure to identify the higher notes can be attributed to the lack of partials detected at high frequencies. Performance tends to tail off towards the end of the piece, due mainly to the increasing number of missing partials.

To summarise, this last result highlights some of the deficiencies of the current models and

algorithms.

1. The partial onset detection performance is degraded when a number of partials are coincident. While it still performs satisfactorily at low frequencies, where the partial magnitudes tend to be fairly large, the increased 'jitter' in signals of this type causes the higher partials, which typically have much less energy, to be missed.

2. It is difficult to detect a partial onset when there is already significant energy before the onset at the same frequency. The present system depends on the signal phase characteristics in the region of the onsets to detect them and so does not perform well when these have become corrupted by coincidence or noise.

3. The note identification algorithm relies heavily on partials being detected and it has been seen that this is not always possible. A possible improvement would be to allow the presence of partials to be proposed from the presence of other, already detected, partials. It may then be possible to track the characteristic energy of such partials, confirming their presence, even though their onsets could not be detected.

7.6 Summary

While not perfect, it is hoped that the results presented in this Chapter have proved the suitability of the Multiresolution Fourier Transform as a signal representation for the analysis of polyphonic music and demonstrated that its potential to allow analysis algorithms to treat scale as an additional representation parameter to time and frequency offers numerous advantages over single or multiscale techniques. Although the use of phase information in such analysis has its limitations, it has been shown that it can provide reduced uncertainty in time and frequency estimates, over what can be accomplished with 'magnitude only' processing. Of course, the use of multiple window scales is a necessary precondition for such methods to be effective, since without it the problems of interference

would render the phase information ineffective.

Chapter 8

Conclusions

8.1 Thesis Summary

The aim of the work described in this thesis has been to explore the application of multiresolution signal processing techniques to the analysis of polyphonic music. The motivation for this work came from the recognition of a problem which has limited the success of previous studies in this area, which were based on STFT signal representations, namely the choice of a suitable analysis window size e.g. [Wat86, Ser89].

8.1.1 Signal Representation

Chapter 2 contains a review of several signal representations which have been applied to audio signals and identified the problem that none of them achieved sufficient resolution in both time and frequency while retaining a structure well suited to the analysis task. Further discussion revealed that no single window function could exist, due to the uncertainty principle, which could satisfy the requirements of such an analysis system. Additionally, even if given some flexibility in the choice of window size, it would not be possible to select the most appropriate size until the analysis was partly complete. The second half of Chapter 2 described a signal representation, the *Multiresolution*

Fourier Transform, which attempts to overcome these difficulties by offering a range of resolutions at all points on the time frequency plane. Analysis algorithms may then be defined which pick the most appropriate coefficients from the range of available resolutions according to some selection criterion. It was shown that a feature's context, the relative positions of neighbouring features, in addition to the known properties of that feature, provides a sufficient set of parameters to determine an appropriate analysis scale.

The implementation details of the MFT were discussed in Chapter 3. A modification of the initial MFT definition, termed relaxation, involving oversampling by a factor of two, was described which resulted in several desirable improvements of the MFT structure. The implementation of both the forward and inverse transforms was discussed, as was the incorporation of the initial fixed length algorithms into a 'blocked' scheme for the analysis of arbitrary duration signals. It was shown that the MFT can be efficiently implemented using FFTs and that the computational cost is comparable with generating the corresponding number of STFTs (Equation 3.17). Finally, the MFT was applied to a set of three simple signals to demonstrate its properties, in particular the way in which the representation of these signals changes with scale.

8.1.2 Signal Modelling and Analysis

Chapter 4 considered the modelling of harmonically based musical signals. The aim was to define a feature hierarchy which could relate musical notes to the kind of signal representation offered by the MFT. At the lowest level of the hierarchy would be the MFT coefficients themselves, with the notes at the top level. It was shown that a direct mapping between these two levels was not possible due to the spectral structure of a note (its partials) being distributed across the time-frequency plane. An intermediate set of features were deduced, all of which could be localised on the time-frequency plane using an appropriate analysis scale. In this way, given the existence of a signal model for a feature, then the known relationship between the corresponding set of coefficients and

the feature's parameters can be used to obtain a description of that feature in which interference from neighbouring features can be minimised by choice of suitable analysis scale.

The set of intermediate features was defined to be:

1. Short segments of note partials in which the amplitude and frequency remains constant.
2. An optional partial onset feature describing the transient behaviour which may be present at the beginning of a partial.

Having established a suitable feature hierarchy, a set of mathematical models for each level of the feature hierarchy were defined. The models incorporated the variations in feature structure common in natural signals by the use of Markov processes.

In Chapter 5, the feature models of Chapter 4 were used to define a set of detection algorithms which enable the detection of each feature from the next lower level in the hierarchy. A common element of these algorithms was the use of sample covariance as a means of dealing with the statistical nature of the signal and improving the reliability of the detection in the presence of noise. In addition, the algorithms made extensive use of the coefficient phase as well as the magnitude. The majority of previous workers have not made use of the phase information available in the signal representations their systems were based on, e.g. [PG78, Ser89, Wat86]. This was done because it was considered worthwhile to try to use this information, both to verify that the MFT presents it in a suitable form and to investigate its usefulness. The analysis of the Schubert trio in Section 7.5, however, showed that such phase information can be corrupted in some circumstances, e.g. when partials from two or more sources are coincident in frequency. In the case of the Schubert analysis, this caused the partial tracking to produce poor frequency estimates, appearing as jitter in the partial tracks and caused the onset detection to fail at even moderate partial magnitudes. In the cases where the expected phase behaviour does exist, the detectors perform well, certainly better than could be expected of an algorithm operating on coefficient magnitude alone. All of the detection algorithms used in this thesis could be easily modified to make them phase independent, by using

only magnitude information. The underlying model structure and estimation algorithms would remain unchanged. The modifications required would simply require the removal of the complex elements from the detection algorithms. The tracking of partials would use the interpolation of local coefficient magnitudes rather than the forward phase differences and the onset detection would substitute the time derivative of $c_{0ik}(n)$, $c'_{0ik}(n)$, for $c'_{1ik}(n)$ (from eqn. 5.27) in its calculation. Similarly, the use of phase in the transient onset detection could be eliminated by just using the local energy at the onset, given by the denominator of Equation 5.43. An obvious strategy for incorporating these modified forms would be to run them in parallel with the originals and then compare their outputs. In this way, the system would be able to operate well with phase coherent and incoherent signals but also be able to determine whether a feature hypothesis was phase coherent or not and use this information to decide whether more than one feature was present at that place. In the case of partials, such a modification would affect the feature hierarchy, introducing a new element (a multi-partial) intermediate between partials and MFT coefficients to allow more than one partial to be present at the same place on the time frequency plane. Several partials could then map onto one multi-partial. The modified feature hierarchy is shown in Figure 8.1. All partials attached to the same multi-partial would have the same path vector parameters during their overlapping portions and it would not be possible to distribute the coefficient energy amongst the partials without having some model of the spectra of the notes to which they are associated. The multi-partial could possibly attempt to unravel the beating present in the MFT coefficients for such cases, using techniques such as those suggested in [Mah90] and [Cha86], given that the new structure makes available an estimate of the number of partials overlapping and allows access to the parameters of their related notes and harmonics.

In was noted in Chapter 7 (two piano notes) that the event detection algorithms fail to detect the upper harmonics of the notes due to their relatively small magnitude. It may be possible to improve this by adapting the detection normalisation process with respect to frequency, given the general

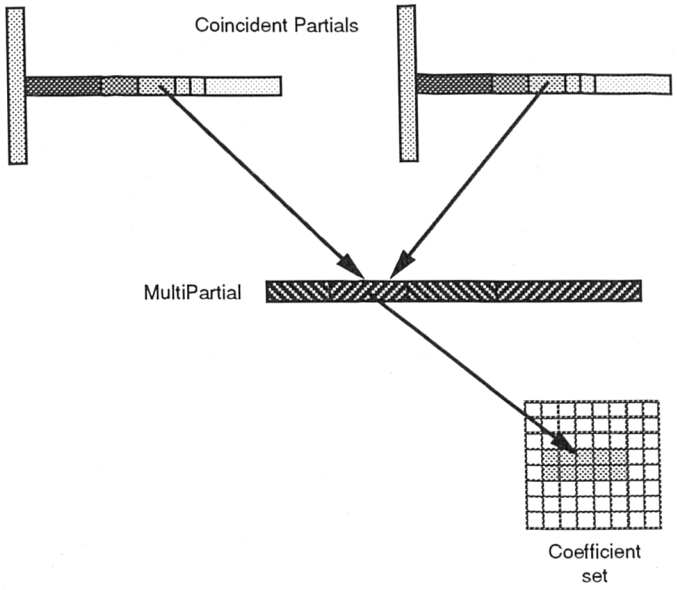

Figure 8.1: Proposed modification to Feature Hierarchy

low-pass frequency spectrum of most music. The disadvantage of such a technique would certainly be an increase in the number of false detections at these frequencies.

8.1.3 Transcription

The algorithms used selectively to apply the feature detectors in a multiresolution framework and combine their outputs across scale were described in Chapter 6. The use of several signal descriptions inevitably introduces complexity into this task, which is not encountered in single resolution systems. In particular, it has been found difficult to keep information regarding the signal context, positions of existing features etc, up to date while attempting to detect all features present in the signal. Very careful ordering of the analysis steps is required and it has been found difficult to avoid introducing significant time lag between the time at which new features are detected and that at which the state of existing features is updated. This results in the suitable analysis scale range for some feature being determined by the state of features some way before that point. For instance, the detection of short duration notes in close frequency proximity, such as trills, remains

a difficulty since the feature context time-lag causes the analysis scale for those note onsets to be unnecessarily restricted in temporal resolution. There is, however, no fundamental reason why this problem cannot be overcome by a more sophisticated algorithm; it is simply a deficiency in the current implementation.

A general criticism that may be made of the current algorithm is that there is not enough information flow between its various parts. The majority of information passes from the lower to the higher levels of the feature hierarchy and elements on each level behave fairly independently. For instance, note hypotheses depend exclusively on detected partials: there is no way for the presence of a partial to be proposed given the existence of one or more note hypotheses or, for that matter, other partials. Also, the partial tracking algorithm could be improved by taking into account the frequency estimates of other partials associated with the same note.

It has been stated that the output of the current system should be suitable as input to subsequent processing stages which may well use culturally specific and other a priori information to further refine the information flow. It would, of course, be highly desirable to pass information relating to the musical context back down from such stages into the feature detection process. The independent model of note probabilities (eqns. 4.5 & 4.6) is weak but necessary when only the lower levels of such a more complete system are being constructed. When available, high-level information, such as 'key' and 'rhythm' could be used to direct the low-level detectors, enabling them to perform more reliably, particularly at low signal to noise ratios.

The current system makes no attempt to account for the energy in the signal. It should be possible to calculate the energy that each detected feature represents and thereby decide whether all significant features had been extracted. This approach was one of the strengths of Charles Watson's work [Wat86] which continued to distribute harmonic energy among note hypotheses until the energy remaining in the signal was very small. Problems of introducing such a scheme into the present system include the fact that general signals are expected to contain significant energy in

forms for which suitable models have not yet been defined and the use of many resolutions requires the accounting to be performed over a wide range of coefficient geometries.

While there is still a great deal of work to be done using the MFT as an analysis tool, its application to sound synthesis has not been investigated in this work. The similarity of the MFT's structure to the model used by granular synthesis is clear. It has been noted [Roa85] that there has been some debate in this area on the choice of suitable grain size. Using the MFT for such work would allow for a range of grain sizes in the synthesis algorithms, allowing for an adaptive, rather than uniform, quantisation of the time-frequency 'sound maps' commonly used to drive such systems.

Less directly, the MFT is a very powerful general analysis and visualisation tool in its own right, which could be used to verify the performance of other synthesis techniques.

8.2 Concluding Remarks

This thesis has presented a new approach to the analysis of musical signals. It has investigated the application of a new signal transform, the Multiresolution Fourier Transform, to the transcription of polyphonic music. The MFT provides a multiplicity of views of a signal, each with differing time-frequency resolutions, in a regular structure enabling the development of analysis algorithms which operate in a space with axes of time, frequency *and* scale. The motivation for this approach was provided by observing that it is not possible to define a single transform analysis window suited to the detection of the wide range of features present in musical signals.

Algorithms have been presented to detect and track the partials of musical notes based on a novel statistical model, and to find groups of these partials to form a list of note estimates. A key part of the analysis involved the confirmation of feature estimates from various scales within the MFT. This process facilitates the selection of suitable analysis scales for each feature as well as increasing the probability of reliable detection and decreasing the uncertainties of the associated parameter estimates.

While the list of note features detected from the signals is far from complete, it is hoped that the results presented demonstrate the success of the technique and that the discussion of the general principles involved points towards similar success for a broader range of features given the flexibility of the MFT as a transform for signal analysis.

The key elements of this work form part of a paper [WCPon] which also includes the application of the MFT in image processing as well as a more rigorous theoretical development of the transform itself. This work was presented at the International Computer Music Conference in Ohio (1989) and Glasgow [PW90] (1990).

Bibliography

[ACE88] J.M. Adrien, R. Caussé, and E.Ducasse. Sound synthesis by physical models. In *Proceedings of the 84th Audio Engineering Society Convention*, 1988.

[ACR87] Jean Marie Adrien, René Causse, and Xavier Rodet. Sound synthesis by physical models, application to strings. In *Proceedings International Computer Music Conference*, 1987.

[And71] T. W. Anderson. *The Statistical Analysis of Time Series*. John Wiley and Sons, 1971.

[AR85] Jean Marie Adrien and Xavier Rodet. Physical models of instruments, a modular approach, application to strings. In *Proceedings International Computer Music Conference*, 1985.

[AS87] E. H. Adelson and E. Simoncelli. Orthogonal pyramid transforms for image coding. In *Proceedings SPIE conference on Visual Communications and Image Processing*, 1987.

[Bas81] M. J. Bastiaans. A sampling theorem for the complex spectrogram and gabor expansion of a signal into gaussian elementary signals. *Opt. Eng.*, 20:594–598, 1981.

[Bla65] E. Donnel Blackham. The physics of the piano. *Scientific American*, pages 88–99, 1965.

[Bra86] R. N. Bracewell. *The Fourier Transform and its Applications. 2nd ed.* McGraw Hill, 1986.

[Cal89] Andrew Calway. *The Multiresolution Fourier Transform: A general Purpose Tool for Image Analysis.* PhD thesis, Department of Computer Science, The University of Warwick, UK, September 1989.

[Can80] Richard Cann. An analysis synthesis tutorial: I-III. *Computer Music Journal*, 3-4, 1979-1980.

[Cha75] C. Chatfield. *The Analysis of Time Series: An Introduction.* J.W.Arrowsmith Ltd, 1975.

[Cha86] Chris Chafe. Source separation and note identification in polyphonic music. Technical Report STAN-M-34, Stanford University, Department of Music, April 1986.

[CJK+85] Chris Chafe, David Jaffe, Kyle Kashima, Bernard Mont-Reynaud, and Julius Smith. Techniques for note identification in polyphonic music. In *Proceedings International Computer Music Conference*, 1985.

[CM80a] T.A.C.M. Classen and W.F.G. Mecklenbräuker. The wigner distribution - a tool for time frequency analysis part I: Continuous-time signals. *Phillips Journal of Research*, 35:217–250, 1980.

[CM80b] T.A.C.M. Classen and W.F.G. Mecklenbräuker. The wigner distribution - a tool for time frequency analysis part II: Discrete-time signals. *Phillips Journal of Research*, 35:276–300, 1980.

[CM80c] T.A.C.M. Classen and W.F.G. Mecklenbräuker. The wigner distribution - a tool for time frequency analysis part III: Relations with other time-frequency signal transformations. *Phillips Journal of Research*, 35:372–389, 1980.

[CM83] T.A.C.M. Classen and W.F.G. Mecklenbräuker. The aliasing problems in the discrete-time wigner distributions. *IEEE Transactions on Acoustics Speech and Signal Processing*, 31:1067–1072, 1983.

[CMR82] Chris Chafe and Bernard Mont-Reynaud. Toward an intelligent editor of digital audio: Recognition of musical constructs. *Computer Music Journal*, 6(1), 1982.

[Coh89] Leon Cohen. Time frequency distributions. *Proceedings of The Institute of Electrical and Electronic Engineers*, 77:941–981, 1989.

[Dau88a] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, XLI:909–996, 1988.

[Dau88b] J. G. Daugman. Complete discrete 2-d gabor transforms by neural networks for image analysis and compression. *IEEE Transactions on Acoustics Speech and Signal Processing*, 36:1169–1179, 1988.

[ECG76] D. Esteban, A. Crosier, and C. Galand. Perfect channel splitting by the use of interpolation/decimation/tree decomposition techniques. In *Proceedings International Conference on Information Science and Systems*, 1976.

[F.74] Kaiser J. F. Nonrecursive digital filter design using the $i_0$sinh window function. *Proceedings of the IEEE International Symposium on Circuits and Systems*, pages 20–23, 1974.

[Gab46] D. Gabor. Theory of communication. *Proceedings IEE*, 1946.

[Gel81] Stanley A. Gelfand. *Hearing*. Marcel Dekker, 1981.

[GG78] J. M. Grey and John W. Gordon. Perception of spectral modifications on orchestral instrument tones. *Computer Music Journal*, 2(1):24–31, 1978.

[GM77] J. M. Grey and James A. Moorer. Perceptual evaluations of synthesised musical instrument tones. *Journal of the Acoustical Society of America*, 62(2):454–462, 1977.

[GM84] A. Grossman and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal of Mathematical Research*, 15:723–736, 1984.

[Gre75] J. M. Grey. An exploration of musical timbre. Technical Report STAN-M-2, Stanford University, Department of Music, 1975.

[Gro89] Numerical Algorithms Group. *NAG Fortran Manual.* NAG, 1989.

[Har78] F. J. Harris. On the use of windows for harmonic analysis with the discrete fourier transform. *Proceedings of The Institute of Electrical and Electronic Engineers*, 66:51–83, 1978.

[Har83] R. Haralick. Image segmentation survey. In o. Faugeras, editor, *Fundamentals in Computer Vision.* C.U.P, 1983.

[Jai89] A. K. Jain. *Fundamentals of Digital Image Processing.* Prentice-Hall, 1989.

[JK83] Cornelis P. Janse and Arie J. M. Kaizer. Time-frequency distributions of loudspeakers: The application of the wigner distribution. *Journal of the Audio Engineering Society*, 31(4):198–222, 1983.

[KM88] R. Kronland-Martinet. The wavelet transform for analysis, synthesis and processing of speech and music sounds. *Computer Music Journal*, 12(4), 1988.

[KMMG87] R. Kronland-Martinet, J. Morlet, and A. Grossmann. Analysis of sound patterns through wavelet transforms. *International Journal of Pattern Recognition and Artificial Intelligence*, 2:97–126, 1987.

[Mah90] Robert C. Maher. Evaluation of a method for separating digitized duet signals. *Journal of the Audio Engineering Society*, 38:956–979, 1990.

[Mal89] S.G. Mallat. A theory for multiresolution signal representation: The wavelet representation. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 11:674–693, 1989.

[Mar82] D. Marr. *Vision.* Freeman, 1982.

[MB79] Stephen McAdams and Albert Bregman. Hearing musical streams. *Computer Music Journal*, 3(4), 1979.

[Moo75] James A. Moorer. On the segmentation and analysis of continuous musical sound by digital computer. Technical Report STAN-M-3, Stanford University, Department of Music, 1975.

[Moo78] James A. Moorer. The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society*, 26:42–45, 1978.

[MQ86] R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics Speech and Signal Processing*, 34:744–754, 1986.

[Nor70] J. O. Nordmark. Time and frequency analysis. In J. V. Tobias, editor, *Foundations of Modern Auditory Theory.* Academic Press, 1970.

[OW87] Frank Opolko and Joel Wapnick. *McGill University Master Samples.* McGill University, 1987.

[Pap77] Athanasios Papoulis. *Signal Analysis.* McGraw-Hill, 1977.

[Pap84] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes.* McGraw-Hill, 1984.

[PG78] Martin Piszczalski and Bernard A. Galler. A method for computer analysis and transcription of musical sound. In *Proceedings of the Research Symposium on the Psychology and Acoustics of Music*, pages 167–189, 1978.

[Por76] Michael R. Portnoff. Implementation of the phase vocoder using the fast fourier transform. *IEEE Transactions on Acoustics Speech and Signal Processing*, 24(3):243–248, 1976.

[Por80] Michael R. Portnoff. Time-frequency representation of digital signals and systems based on short-time fourier analysis. *IEEE Transactions on Acoustics Speech and Signal Processing*, 28(1):55–69, 1980.

[Por81] Michael R. Portnoff. Short-time Fourier analysis of sampled speech. *IEEE Transactions on Acoustics Speech and Signal Processing*, 29(3):364–373, 1981.

[PSL61] H. O. Pollak, D. Slepian, and H. J. Landau. Prolate spheroidal wave functions, fourier analysis and uncertainty I-III. *BSTJ 40-41*, 1961.

[PW90] Edward R.S. Pearson and R. G. Wilson. Musical event detection from audio signals within a multiresolution framework. In *Proceedings International Computer Music Conference*, 1990.

[RDP87] X. Rodet, P. Depalle, and G. Poirot. Speech analysis and synthesis methods based on spectral envelopes and voiced/unvoiced functions. In *Proceedings of the European Conference on Speech Technology*, 1987.

[RG75] L. R. Rabiner and B. Gold. *The Theory and Application of Digital Signal Processing.* Prentice-Hall, 1975.

[Roa85] Curtis Roads. Granular synthesis of sound. In Curtis Roads and John Strawn, editors, *Foundations of Computer Music*. MIT Press, 1985.

[Rod84] Xavier Rodet. Time-domain formant-wave function synthesis. *CMJ*, 8(3):9–14, 1984.

[RPB84] Xavier Rodet, Y. Potard, and J.B Barrière. The chant project: From synthesis of the singing voice to synthesis in general. *CMJ*, 8(3):15–31, 1984.

[SB87] M. J. T. Smith and T.P. Barnwell. A new filter bank theory for time-freqency representation. *IEEE Transactions on Acoustics Speech and Signal Processing*, 35:314–326, 1987.

[Ser89] Xavier Serra. A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition. Technical Report STAN-M-58, Stanford University, CCRMA, Department of Music, 1989.

[Sle76] David Slepian. On bandwidth. *Proceedings of The Institute of Electrical and Electronic Engineers*, 64:292–300, 1976.

[Sma37] A. M. Small. An objective analysis of artistic violin performance. Technical Report 4, University of Iowa, Studies in Music Perception, 1937.

[Sma70] Arnold M. Small. Periodicity pitch. In J. V. Tobias, editor, *Foundations of Modern Auditory Theory*. Academic Press, 1970.

[Smi87] Julius O. Smith. Efficient simulation of the reed-bore and bow-string mechanisms. In *Proceedings International Computer Music Conference*, 1987.

[SS86] Julius O. Smith and Xavier Serra. Parshl: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation. In *Proceedings International Computer Music Conference*, 1986.

[VKDV88] D. J. Verschuur, A. J. M. Kaizer, W. F. Druyvesteyn, and D. De Vries. Wigner representation of loudspeaker responses in a living room. *Journal of the Audio Engineering Society*, 36(4):203–212, 1988.

[War70] W. Dixon Ward. Musical perception. In J. V. Tobias, editor, *Foundations of Modern Auditory Theory*. Academic Press, 1970.

[Wat86] Charles Watson. *The Computer Analysis of Polyphonic Music.* PhD thesis, The University of Sydney, Australia, 1986.

[WCPon] Roland G. Wilson, Andrew D. Calway, and Edward R.S. Pearson. A generalised wavelet transform for fourier analysis: the multiresolution fourier transform and its application to image and audio signal analysis. *IEEE Transactions on Information Theory*, Accepted for publication.

[WK88] R. Wilson and H. Knutsson. Uncertainty and inference in the visual system. *IEEE Transactions SMC-18*, pages 305–312, 1988.

[WS87a] Roland G. Wilson and M. Spann. Finite prolate spheroidal sequences and their applications I-II. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 1987.

[WS87b] Roland G. Wilson and M. Spann. *Image Segmentation and Uncertainty.* Research Studies Press, 1987.